# THE ANNALS *of* MATHEMATICAL STATISTICS

The Annals of Mathematical Statistics Is Affiliated with the American Statistical Association and Is Devoted to the Theory and Application of Mathematical Statistics

Volume VII, Number 3
SEPTEMBER, 1936

PUBLISHED QUARTERLY
ANN ARBOR, MICHIGAN

*Four Dollars per annum*

*Made in United States of America*

Address: Annals of Mathematical Statistics
Post Office Box 171, Ann Arbor, Michigan

Composed and Printed at the
WAVERLY PRESS, Inc.
Baltimore, Md.

# ON A METHOD OF TESTING THE HYPOTHESIS THAT AN OBSERVED SAMPLE OF $n$ VARIABLES AND OF SIZE $N$ HAS BEEN DRAWN FROM A SPECIFIED POPULATION OF THE SAME NUMBER OF VARIABLES

By John W. Fertig

With the Technical Assistance of Margaret V. Leary*

The problem of determining whether or not a given observation may be regarded as randomly drawn from a certain population completely specified with respect to its parameters is readily solved if the probability integral of that population be known. In particular if the population specified be a normal population, one may calculate the relative deviate $(x - a)/\sigma$, where $a$ and $\sigma$ are the population mean and standard deviation respectively, and refer to tables of the normal probability integral. The hypothesis that $x$ was drawn from this population may be rejected if $P$ is less than an arbitrarily fixed value, say $\leqq .01$. Generalizations of this problem may be made in two directions: 1) May a single observation simultaneously made on $n$ variables be considered as randomly drawn from a specified population of $n$ variables? 2) May a sample of one variable and of size $N$ be regarded in its entirety as randomly drawn from a specified univariate population?

The solution to the first problem for the case of sampling from a normal population of $n$ variables was given by Karl Pearson in 1908[1] as the "Generalized Probable Error." Let

$$\chi^2 = \frac{1}{P}\left\{ \mathop{S}_{i,j=1}^{n} P_{ij} \left[ \frac{(x_i - a_i)(x_j - a_j)}{\sigma_i \sigma_j} \right] \right\}$$

where $a_i$ and $\sigma_i$ are the population mean and standard deviation respectively of the $i^{\text{th}}$ variable, and $P_{ij}$ is the usual cofactor of the element in the $i^{\text{th}}$ row and $j^{\text{th}}$ column of the determinant $P$ of population correlation coefficients. That is,

$$P = |\rho_{ij}|;\ i, j = 1, 2, 3, \cdots, n.$$

The probability of an observation yielding a smaller discrepancy than that represented by the value of $\chi^2$, i.e., lying between 0 and $\chi^2$, may then be calculated from Tables of the Incomplete Normal Moment Functions[2]. The tables are entered in terms of $(\chi^2)^{\frac{1}{2}}$ and $(n - 1)$, and the tabled value multiplied by $(2\ \pi)^{\frac{1}{2}}$ or 2 depending upon whether $n$ be even or odd respectively.

* From the Memorial Foundation for Neuro-Endocrine Research and the Research Service of the Worcester State Hospital, Worcester, Massachusetts.

The probability of an observation giving a greater discrepancy is then the complement of this value. Obviously, this latter probability may be obtained directly by entering tables of the $X^2$ distribution such as Elderton's[3] with $n$ degrees of freedom, or through the use of Tables of the Incomplete Γ-Function[4].

The second problem, limited to the case of sampling from a normal population, was investigated by J. Neyman and E. S. Pearson in 1928[5]. The observed sample may be regarded as a point in $N$-dimensional space, where $N$ is the sample size. Criteria for the acceptance or rejection of the hypothesis may be associated with contour surfaces in this space, so that in moving out from contour to contour the hypothesis becomes less and less reasonable. Frequently, contour surfaces on which the mean or standard deviation is constant are used for the testing of this hypothesis. Such surfaces are deficient inasmuch as they are not "closed" contours. Another contour system which appears more satisfactory is that of equiprobable pairs of $m$ and $s$. The latter system in fact encloses roughly the same region as do the separate contours for the means and standard deviations. These systems are of course dependent on the particular statistics chosen to describe the sample and are further limited in that they do not take into account the probability of alternative hypotheses concerning the origin of the sample.

Using the principle of maximum likelihood Neyman and Pearson have developed a system of contours which is free of the above limitations. The system so derived is in fact quite similar to that of equiprobable pairs $m$ and $s$. In a later paper[6], these same investigators have shown that this method of maximum likelihood does enable one to select the most efficient criteria for the testing of an hypothesis. The criterion selected on this basis is defined as

$$\lambda = \frac{\text{Likelihood that sample came from specified population}}{\text{Maximum likelihood that sample came from some other population}}$$

$$= (s^2/\sigma^2)^{N/2} e^{-N/2} \left[ \frac{s^2 + (\bar{x} - a)^2}{\sigma^2} - 1 \right]$$

where $a$ and $\sigma$ are the population mean and standard deviation respectively, and $\bar{x}$ and $s$ the sample mean and standard deviation.

$\lambda$ is constant upon certain contour surfaces in $N$-dimensional space, and diminishes on passing outward. The form of the surfaces is independent of $N$. It is evident that $\lambda$ must lie between zero and unity. When it is close to unity we know that it is reasonable to assume that our hypothesis is true, when small we know that it is unreasonable. But we must know the probability of $\lambda$ less than a certain value occurring when the hypothesis tested is true, so that we may control another source of error, namely, that of rejecting the hypothesis when it is true. In other words, we must know the sampling distribution of $\lambda$, so that we will reject the hypothesis only when the probability of obtaining a smaller value is negligible, say $P_\lambda \leqq .01$. Neyman and Pearson were not able to evaluate this distribution but they were able to integrate the original density function of the population appropriate to $N$-dimensional space outside of the various $\lambda$ contours. This they were able to do by effecting a transformation of the density function and contours to the plane of $m$ and $s$. These values of $P_\lambda$ have been tabled by them[7], the tables being entered in terms of $N$ and $k$, where

$$k = \log \left[\frac{s^2 + (\bar{x} - a)^2}{\sigma^2}\right] - \log (s^2/\sigma^2)$$

The generalization of either of the above problems requires a criterion to test an hypothesis which may be formulated as follows: Given a sample $\Sigma$ of $n$ variables and of size $N$ with means $\bar{x}_1, \bar{x}_2, \cdots, \bar{x}_n$, standard deviations $s_1, s_2, \cdots, s_n$, and correlation coefficients $r_{12}, r_{13}, \cdots, r_{1n}, r_{23}, \cdots, r_{2n}, \cdots, r_{(n-1)n}$, may we regard this sample as randomly drawn from a population $\pi$ of $n$ variables and completely specified with respect to all its parameters? We shall restrict our inquiries to the case where $\pi$ is a normal population. In this case the distribution law is

$$f(x_1, x_2, \cdots, x_n) = \frac{1}{(2\pi)^{n/2}\sigma_1\sigma_2 \cdots \sigma_n P^{\frac{1}{2}}} e^{-\phi}$$

where

$$\phi = +\frac{1}{2P}\left\{ \underset{i,j=1}{\overset{n}{S}} P_{ij}\left[\frac{(x_i - a_i)(x_j - a_j)}{\sigma_i\sigma_j}\right]\right\}$$

where $a_i$ and $\sigma_i$ are the population mean and standard deviation respectively of the $i^{\text{th}}$ variable, and $P$ and $P_{ij}$ are as previously defined.

Thus the probability that $\Sigma$ has been drawn from $\pi$ with its $N$ values of $x_{i\alpha}$ $(i = 1, 2, \cdots, n)$ lying in the interval $x_{i\alpha} \pm \frac{1}{2}dx_{i\alpha}$; $(\alpha = 1, 2, \cdots, N)$ is given by

$$C = \left[\frac{1}{(2\pi)^{n/2}\sigma_1\sigma_2 \cdots \sigma_n P^{\frac{1}{2}}}\right]^N e^{-\Theta}\, dX$$

where

$$\begin{aligned}\Theta &= \frac{1}{2P}\left\{ \underset{i,j=1}{\overset{n}{S}} P_{ij} \underset{\alpha=1}{\overset{N}{S}} \left[\frac{(x_{i\alpha} - a_i)(x_{j\alpha} - a_j)}{\sigma_i\sigma_j}\right]\right\} \\ &= \frac{N}{2P}\left\{ \underset{i,j=1}{\overset{n}{S}} P_{ij}\left[\frac{s_i s_j r_{ij} + (\bar{x}_i - a_i)(\bar{x}_j - a_j)}{\sigma_i\sigma_j}\right]\right\}\end{aligned}$$

$$dX = \prod_{i=1}^{n}\prod_{\alpha=1}^{N} dx_{i\alpha}$$

The likelihood that $\Sigma$ has been drawn from any other normal population, such as $\pi'$, is given by

$$C' = \left[\frac{1}{(2\pi)^{n/2}\sigma_1'\sigma_2' \cdots \sigma_n' P'^{\frac{1}{2}}}\right]^N e^{-\Theta'}\, dX$$

where

$$\Theta' = \frac{N}{2P'}\left\{\underset{i,j=1}{\overset{n}{S}} P'_{ij}\left[\frac{s_i s_j r_{ij} + (\bar{x}_i - a'_i)(\bar{x}_j - a'_j)}{\sigma'_i \sigma'_j}\right]\right\}$$

The population from which it is most likely that $\Sigma$ has been drawn is that for which $C'$ is a maximum. The values of the parameters of this population may be obtained by putting

$$\frac{\partial C'}{\partial a'_i} = 0, \qquad \frac{\partial C'}{\partial \sigma'_i} = 0; \qquad (i = 1, 2, \cdots, n)$$

$$\frac{\partial C'}{\partial \rho'_{ij}} = 0; \qquad (i, j = 1, 2, \cdots, n)$$

These conditions are fulfilled when

$$a'_i = \bar{x}_i;\ \sigma'_i = s_i; \quad (i = 1, 2, \cdots, n)$$

$$\rho'_{ij} = r_{ij}; \quad (i, j = 1, 2, \cdots, n)$$

So that

$$C'_{\text{max.}} = \left[\frac{1}{(2\pi)^{n/2} s_1 s_2 \cdots s_n R^{\frac{1}{2}}}\right]^N e^{-nN/2}$$

where

$$R = |r_{ij}|; \qquad i, j = 1, 2, \cdots, n$$

The appropriate criterion to select in order to test our hypothesis is thus

$$\lambda = \frac{C}{C'_{\text{max.}}} = \left[\frac{s_1 s_2 \cdots s_n R^{\frac{1}{2}}}{\sigma_1 \sigma_2 \cdots \sigma_n P^{\frac{1}{2}}}\right]^N e^{-w}$$

where

$$w = \frac{N}{2}\left\{\underset{i,j=1}{\overset{n}{S}} \frac{P_{ij}}{P}\left[\frac{s_i s_j r_{ij} + (\bar{x}_i - a_i)(\bar{x}_j - a_j)}{\sigma_i \sigma_j}\right] - n\right\}$$

The equations $\lambda$ = constant represent a series of contours in $N$-dimensional space. As we move outward from contour to contour our hypothesis becomes less and less acceptable. Although we may be confident that the use of this criterion will minimize the chance of accepting the hypothesis when it is false we must know the frequency with which samples occur outside of a given $\lambda$ contour when the hypothesis is true. In other words, we must know the integral of $C$ outside of various contours, or else we must know the sampling distribution of $\lambda$. The former is an exceedingly difficult method for $n$ greater than unity. Thus for the case of $n = 2$ we should have to integrate some such expression as

$$k s_1^{N-2} s_2^{N-2} e^{-\Theta}(1 - r_{12}^2)^{\frac{N-4}{2}}\, d\bar{x}_1 d\bar{x}_2 ds_1 ds_2 dr_{12}$$

outside of the various contours. Nor have we so far been able to evaluate the sampling distribution. We can however give an expression for the moments of $\lambda$ and thus reach an approximate distribution.

Wilks[8] has derived expressions for the moment coefficients about zero for the maximum likelihood criterion that $k$ samples of $n$ variables and of $N_t$ observations each have been drawn from the same unspecified normal population of $n$ variables. Thus,

$$\mu_h'(\lambda) = \prod_{t=1}^{k} \left\{ \left[ \frac{\overset{k}{\underset{t=1}{S}} N_t}{N_t} \right]^{\frac{h\,n\,N_t}{2}} \prod_{i=1}^{n} \left[ \frac{\Gamma\left(\frac{N_t(1+h)-i}{2}\right)}{\Gamma\left(\frac{N_t-i}{2}\right)} \right] \right\}$$

$$\prod_{i=1}^{n} \left\{ \frac{\Gamma\left[\frac{\overset{k}{\underset{t=1}{S}} N_t - i}{2}\right]}{\Gamma\left[\frac{(1+h)\overset{k}{\underset{t=1}{S}} N_t - i}{2}\right]} \right\}$$

from which we can write expressions giving the moment coefficients about zero for the $\lambda$ criterion for two samples

$$\mu_h'(\lambda) = \frac{(N_1+N_2)^{\frac{n\,h\,(N_1+N_2)}{2}}}{N_1^{\frac{n\,h\,N_1}{2}}\, N_2^{\frac{n\,h\,N_2}{2}}}$$

$$\prod_{i=1}^{n} \left\{ \frac{\Gamma\left[\frac{N_1(1+h)-i}{2}\right] \Gamma\left[\frac{N_2(1+h)-i}{2}\right] \Gamma\left[\frac{N_1+N_2-i}{2}\right]}{\Gamma\left(\frac{N_1-i}{2}\right) \Gamma\left(\frac{N_2-i}{2}\right) \Gamma\left[\frac{(N_1+N_2)(1+h)-i}{2}\right]} \right\}$$

The limit of this latter expression as $N_2 \to \infty$ will be the moment coefficient about zero for the $\lambda$ criterion that one sample has been drawn from a specified population. Thus

$$\underset{N_2 \to \infty}{\text{Lim.}}\ \mu_h'(\lambda) = \prod_{i=1}^{n} \left\{ \frac{\Gamma\left[\frac{N_1(1+h)-i}{2}\right]}{\Gamma\left(\frac{N_1-i}{2}\right)} \right\} \left(\frac{2e}{N_1}\right)^{\frac{n\,h\,N_1}{2}} (1+h)^{\frac{-nN_1(1+h)}{2}}$$

Various roots of $\lambda$ are distributed to a good degree of approximation according to a function of the form

$$f(t) = \frac{\Gamma(m_1+m_2)}{\Gamma(m_1)\,\Gamma(m_2)}\, t^{m_1-1}\,(1-t)^{m_2-1}$$

where

$$m_1 = \mu_1'(\mu_1' - \mu_2')/(\mu_2' - \mu_1'^2) \, ; \qquad m_2 = (1 - \mu_1')m_1/\mu_1'$$

and the value of $\mu_h'$ for roots of λ may be obtained by replacing $h$ in the original expression by $h$ times the desired root. Measures of the skewness and kurtosis of this distribution are given by

$$B_1 = 4(m_1 - m_2)^2(m_1 + m_2 + 1)/m_1m_2(m_1 + m_2 + 2)^2$$

$$B_2 = 3B_1(m_1 + m_2 + 2) + 6(m_1 + m_2 + 1)/2(m_1 + m_2 + 3)$$

A comparison with the true measures of skewness and kurtosis for various roots of λ as given by

$$B_1 = \mu_3^2/\mu_2^3 \, ; \qquad B_2 = \mu_4/\mu_2^2$$

will afford a measure of the goodness of the approximation and the range of values of $N$ for which any particular root will be distributed as assumed.

Investigating the moments for $n$ from one to four and $N$ from three to fifty we note that in the case of samples of two and three variables, $\lambda^{1/N}$ follows the assumed distribution for $N$ from 3 to 15; $\lambda^{2/N}$ from 15 to 30; $\lambda^{3/N}$ from 30 to 50. In the case of four variables, $\lambda^{1/2N}$ follows the distribution for $N$ from 5 to 10; $\lambda^{1/N}$ from 10 to 20; $\lambda^{2/N}$ from 20 to 40; $\lambda^{3/N}$ from 40 to 50. It appears likely that for higher values of $n$, for $N$ small, some such root as $\lambda^{1/2N}$ or $\lambda^{1/3N}$ will follow the assumed distribution, while as $N$ increases smaller roots will follow it. For any value of $n$, the smallest permissible value of $N$ is $(n + 1)$.

The probability that a smaller value of λ will be obtained when the sample has actually been drawn from $\pi$, i.e., $P_\lambda$, may thus be obtained by reference to Tables of the Incomplete $B$-Function[9] with $p = m_1$, $q = m_2$, $x =$ value of the particular root of the observed λ. We may also get the 1% and 5% levels of significance directly from Fisher's[10] tables of "$z$" or Snedecor's[11] tables of "$F$" $(= e^{2z})$, by taking

$$n_1 = 2m_2 \, ; \qquad n_2 = 2m_1 \, ; \qquad L = n_2/(n_2 + n_1F) \, ,$$

where $L$ is the desired root of λ. Linear interpolation will generally suffice except for very small values of $N$.

For the case of $N \rightarrow \infty$, we háve

$$\lim_{N \to \infty} \mu_h'(\lambda) = (1 + h)^{-\overset{n+1}{\underset{i=2}{S}} i/2}$$

Thus the quantity $(-2 \log \lambda)$ will be distributed in the $\chi^2$ distribution with $\overset{n}{\underset{i=2}{S}} i$ degrees of freedom.

A table of the 1% and 5% levels of significance for $n$ equal one to four, and values of $N$ from five to $\infty$ is given below

*5% and 1% Levels of Significance of "λ"*

— N —

| $n$ | | 5 | 10 | 15 | 20 | 30 | 40 | 50 | $\infty$ |
|---|---|---|---|---|---|---|---|---|---|
| 1 | 5% | .025 | .037 | .041 | .043 | .045 | .046 | .047 | .050 |
| | 1% | .003 | .006 | .008 | .008 | .009 | .009 | .009 | .010 |
| 2 | 5% × $10^{-2}$ | .046 | .173 | .234 | .269 | .308 | .330 | .343 | .392 |
| | 1% × $10^{-3}$ | .026 | .168 | .260 | .305 | .372 | .409 | .428 | .525 |
| 3 | 5% × $10^{-3}$ | .001 | .036 | .072 | .097 | .125 | .143 | .155 | .211 |
| | 1% × $10^{-4}$ | $.000^{+}$ | .019 | .047 | .076 | .101 | .117 | .128 | .194 |
| 4 | 5% × $10^{-5}$ | | .026 | .106 | .174 | .295 | .356 | .418 | .710 |
| | 1% × $10^{-6}$ | | .007 | .040 | .075 | .145 | .185 | .221 | .466 |

A check on the accuracy of the method of approximation used may be obtained by comparing the values of $P_\lambda$ for the case of $n = 1$ with the exact values given by Neyman and Pearson. For $n = 10$, $\lambda^{1/N}$ is distributed as assumed with $m_1 = 9.0562$, $m_2 = 0.9987$. For the case of $(\bar{x} - a)/\sigma = 0.2$, $s/\sigma = 1.2$, we find $k = 0.48439$, $\lambda^{1/N} = .94395$. From the Tables of the Incomplete $B$-Function we find $P_\lambda = .5936$, from Neyman and Pearson's tables, .5935.

No studies have been made on the extent of deviation from normality permissible for the application of the test. There is no reason to doubt, however, that as much deviation is permissible as in the case of the univariate λ. From theoretical considerations and from sampling studies Neyman and Pearson conclude that the univariate λ technique holds for deviation from normality to the extent of ±0.5 for $B_1$ and 2.5 to 4.2 for $B_2$.

We are confident that this generalized λ technique will be found useful in biological research. If the $n$ variables were uncorrelated we would be able to test whether the sample had been drawn from the population of $n$ variables by successive applications of the univariate λ technique and then combining the resulting probabilities. In general, however, there will be some correlation between the variables, however slight. The method here proposed will take account of all possible intercorrelations, and consequently all multiple and partial correlations.

Now, if $P_\lambda$ is less than some arbitrarily fixed value, say $\leqq .01$, we may decide which variable or variables contributes most to this result, by performing simpler λ tests. It may be due to one or more of the means, standard deviations, or correlation coefficients. As may often be the case, it is not due to any one factor but to contributions from all of them. That is, all possible factors tested separately might show a fairly reasonable value of $P$, but if all the separate values are combined somehow, as by means of this $\lambda$ method, the resultant $P$ may be too small. It is in such problems that this technique should provide valuable information.

In case $k$ samples of $n$ variables are available it should be possible to determine whether all of them have come from the same specified population of $n$ variables by performing $k$ $\lambda$ tests and combining the separate values of $P_\lambda$. Such a hypothesis may best be tested, however, by a further extension of the $\lambda$ theory which the writers are at present investigating.

The following problem is chosen to illustrate the computations involved in the application of the test. Many of the investigations pursued at the Worcester State Hospital attempt to differentiate between schizophrenic patients and normal controls. In one such type of investigation various blood constituents were determined, namely, Urea $N_2$ (mg./100 cc.), Uric Acid $N_2$ (mg./100 cc.), Creatine $N_2$ (mg./100 cc.) for a sample of twenty-five schizophrenic patients. Previous investigations on these same variables for a large series of normal controls yielded constants which for the purpose of the example may be considered as the population parameters. Past studies on these variables have not shown any marked degree of non-normality for the various distributions.

These variables are designated as

$$1 = \text{Urea } N_2\,; \qquad 2 = \text{Uric Acid } N_2\,; \qquad 3 = \text{Creatine } N_2$$

The parameters of the population are given by

$$a_1 = 16.03\,; \qquad a_2 = 1.40\,; \qquad a_3 = 1.25$$

$$\sigma_1^2 = 20.268\,; \qquad \sigma_2^2 = 0.029\,; \qquad \sigma_3^2 = 0.025$$

$$\rho_{12} = .3075\,; \qquad \rho_{13} = .1232\,; \qquad \rho_{23} = .3853$$

The statistics for the sample of twenty-five are

$$\bar{x}_1 = 15.56\,; \qquad \bar{x}_2 = 1.42\,; \qquad \bar{x}_3 = 1.25$$

$$s_1^2 = 10.486\,; \qquad s_2^2 = 0.043\,; \qquad s_3^2 = 0.025$$

$$r_{12} = -.0161\,; \qquad r_{13} = .0925\,; \qquad r_{23} = .2174$$

None of these statistics differs significantly from the corresponding parameters.

$$R = 0.9443\,; \qquad P = 0.7710\,;$$

$$P_{12}/P = -0.3373\,; \qquad P_{13}/P = -0.0061\,; \qquad P_{23}/P = -0.4506\,;$$

$$P_{11}/P = 1.1045\,; \qquad P_{22}/P = 1.2773\,; \qquad P_{33}/P = 1.1744$$

$$w = 12.5\ (0.3802) = 4.7531$$

$$(s_1^2\ s_2^2\ s_3^2\ R/\sigma_1^2\ \sigma_2^2\ \sigma_3^2\ P) = 0.9001$$

$$\log \lambda = 12.5 \log (0.9001) - 4.7531 \log e = \bar{3}.3641$$

$$\lambda = .0023$$

Since the 5% level of significance is about .0001, we thus conclude that the patients are not differentiated from the control population with respect to these variables.

## REFERENCES

1. Pearson, Karl. Biometrika, vol. 6, 1908. pp. 59–68.
2. Tables for Statisticians and Biometricians, Part I. pp. xxiv–xxviii, 22–23.
3. Ibid. pp. xxxi–xxxiii, 26–28.
4. Tables of the Incomplete Γ-Function, 1934.
5. Neyman, J. and Pearson, E. S. Biometrika, vol. 20, 1928. pp. 175–241.
6. Ibid. Phil. Trans. Roy. Soc. A, vol. 231, 1933. pp. 289–337.
7. Tables for Statisticians and Biometricians, Part II. pp. clxxx–clxxxv, 221–223.
8. Wilks, S. S. Biometrika, vol. 24, 1932. pp. 471–494.
9. Tables of the Incomplete Beta-Function, 1934.
10. Fisher, R. A. Statistical Methods for Research Workers. Fourth Edition, 1932.
11. Snedecor, G. W. Calculation and Interpretation of Analysis of Variance and Covariance, 1934.

# ON CONFIDENCE RANGES FOR THE MEDIAN AND OTHER EXPECTATION DISTRIBUTIONS FOR POPULATIONS OF UNKNOWN DISTRIBUTION FORM

By William R. Thompson

About the commonest situation with which we are confronted in mathematical statistics is that where we have a sample of $n$ observations, $\{x_i\}$, which is assumed to have been drawn at random from an unknown population, $U$, with a zero probability that any two values in the finite sample be equal; and we desire to obtain from this evidence some insight as to parameters of the parent population, $U$. If further assumptions are made as to some of the parameters or the form of $U$, there may result a gain in power in testing other given hypotheses or establishing *confidence ranges* for particular parameters, but at an obvious sacrifice of scope in application. Insistent problems involve estimation of mathematical expectation that in further sampling we shall find $x$ lying within a given interval, or similar expectation with regard to parameters of $U$ such as the unknown median. It might seem that, without further assumption, all we should claim is that it is possible to draw from $U$ the sample actually observed. A mere description of the experience may well be considered the observer's first duty, but a restriction to this would leave entirely unused the quality of *randomness* which has been assumed. What additional statements as to $U$ may be appropriate in view of this randomness are our immediate concern; and the object of the present communication is to show how we may obtain such expressions in the form of mathematical expectations, and to present some results. Widespread applications to problems of estimation of *normal ranges of variation* or specific confidence ranges and comparisons of sample reflections of possibly different populations are immediately suggested, and a new foundation is offered for the study of frequency-distribution from the point of view of Schmidt.[1]

## Section 1

Accordingly, consider the following situation. Let $A = \{x\}$ denote the set of all real numbers; and $U$ denote an unknown frequency-distribution law of draft from $\{x\}$ such that there exists an unknown function, $f(x)$, bounded and not negative in $A$, and that the probability of obtaining $x$ in an arbitrary interval $(\alpha, \beta)$ is

$$(1) \qquad P(\alpha < x < \beta) = \int_\alpha^\beta f(x)\cdot dx\,;$$

[1] Schmidt, R., *Annals of Math. Stat.*, *5*, 30, (1934).

and, for every positive $p < 1$, there exists a finite interval $(\alpha, \beta)$ such that $P(\alpha < x < \beta) > p$. Let $U$ be called an *infinite population*; and let $n$ drafts, independently thus governed, made from $A$ *without replacements* be called *a random sample of n observations from U*. Let $S = \{x_k\}$, $k = 1, \cdots, n$, denote such a sample; the enumeration to be made in an arbitrarily determined manner. In any case $x_i \neq x_j$ for $i \neq j$.

Temporarily, let us consider $k$ to indicate the order of draft of the values of $\{x_k\}$, and let $p_k = P(x < x_k)$ denote the probability that $x$, drawn at random from $U$, be less than $x_k$ of $S$. The probability *à priori* (i.e., without regard to relative values of $x$ in the sample) that in such random sampling $p_k$ lie between $p'$ and $p''$, where $0 \leqq p' < p'' \leqq 1$, is obviously independent of $k$, and equals $p'' - p'$; i.e., $p_k$ is equally likely *à priori* to lie in either of any two equal intervals in its possible range, (0, 1). Furthermore, the probability that in the rest of the sample, $S$, there will be just $r$ values less than $x_k$ is

$$\binom{n-1}{r} \cdot p_k^r \cdot (1 - p_k)^{n-1-r},$$

where $r$ is an integer and $0 \leqq r < n$. Of course, $p_k$ is unknown; but we may calculate (for all cases in repeated sampling wherein the same value of $r$ is encountered) the expectation, $\overline{P}_r\ (p' < p_k < p'')$, that $p_k$ lie in the interval $(p', p'')$. This is given by

$$(2) \qquad \overline{P}_r(p' < p_k < p'') = \frac{(r + s + 1)\,!}{r\,!\,s\,!} \int_{p'}^{p''} p^r \cdot q^s \cdot dp\,,$$

where $s = n - 1 - r$, and $q = 1 - p$. This is a familiar result[2, 3, 4] in applications of the well-known principle of Bayes to estimation of *à posteriori* probability. The approach is convenient in that many relations which have been developed in this connection are made immediately available. However, that $p_k$ is equally likely *à priori* to lie in either of any two equal intervals in its possible range, is not based in the present case upon an especially added assumption nor any plea concerning *equal distribution of ignorance,* but follows directly from the elementary assumptions of random sampling. Accordingly, we are enabled to develop for given ranges what may be called the *specific confidence* or mathematical expectation that a given variable lie therein.

Obviously, (2) does not depend on $k$ if this index is the order of draft provided that just $r$ values of the sample, $S$, are less than the one under consideration, $x_k$. To simplify notation, accordingly, let the index $k$ for any given sample, $\{x_k\}$,

[2] Bayes, *Philosophical Transactions, 53,* 370 (1763). *Cf.* Todhunter, I., "A History of the Mathematical Theory of Probability," Macmillan and Co., London, 1865.

[3] Laplace, "Théorie Analytique des Probabilités," Paris, 1820; and other works, *Cf.* Todhunter, *l.c.*

[4] Pearson, K., *Philosophical Magazine, Series 6, Vol. 13,* 365, (1907).

be determined by the relations, $x_i < x_j$ for $i < j$, where $k = 1, \cdots, n$. Then, by (2) as $k = r + 1$, we have

$$(3) \qquad \overline{P}(p' < p_k < p'') = \frac{n!}{(k-1)!\,(n-k)!} \int_{p'}^{p''} p^{k-1} \cdot q^{n-k} \cdot dp\,,$$

where $p_k$ is the probability that random sample values from $U$ will be less than the $k$-th value in order of ascending magnitude from a given random sample, $\{x_k\}$, of $n$ values from $U$; and $\overline{P}(p' < p_k < p'')$ denotes the expectation that in such sampling $p_k$ will lie in the interval, $(p', p'')$.

In general, let $E(w) \equiv \overline{w}$ denote the mathematical expectation of any variable, $w$, under the given sampling conditions. Then, from a well-known relation developed by Laplace, we obtain from (3) the mean expectation of $p_k$,

$$(4) \qquad \bar{p}_k = \frac{k}{n+1}\,;$$

and, further relations[4] of Karl Pearson yield

$$(5) \qquad E((p_k - \bar{p}_k)^2) = \overline{\sigma^2_{p_k}} = \frac{k(n-k+1)}{(n+1)^2 \cdot (n+2)}\,;$$

i.e., the mean squared error in systematic use of $\frac{k}{n+1}$ instead of the unknown $p_k$ should have the value in (5). Specific confidence ranges for $x$ are readily established; e.g., the expectation that in random draft from $U$ we obtain $x$ within the range $(x_k, x_{n-k+1})$ in view of the sample, $S$, is

$$(6) \qquad \overline{P}(x_k < x < x_{n-k+1}) = \frac{n+1-2k}{n+1}, \quad \text{for } 2k < n+1\,;$$

and $\overline{P}(x < x_k) = \overline{P}(x > x_{n-k+1}) = \frac{k}{n+1}$. For a given variate, $w$, the range $(\alpha, \beta)$ will be called *central* if $\overline{P}(w < \alpha) = \overline{P}(w > \beta)$, as in the case under (6). This is in accord with the development of the subject of confidence ranges by Neyman[5,6] and by Clopper and E. S. Pearson[7] following the introduction of the notion of fiducial interval by R. A. Fisher.[8,9] The estimates of $p_k$ in (4) may be of value in studying frequency-distribution from the point of view developed by Schmidt,[1] by comparison of $x_k$ with $\psi\left(\frac{k}{n+1}\right)$ rather than $\psi\left(\frac{2k-1}{2n}\right)$ where $\psi$ is a univariant inverse of the integral of a given frequency function, taken to

[5] Neyman, J., *J. Roy. Stat. Soc.*, *97*, 589, (1934).

[6] Neyman, J., *Annals of Math. Stat.*, *6*, No. 3, 111, (1935).

[7] Clopper, C. J., and Pearson, E. S., *Biometrika*, *26*, 404, (1934).

[8] Fisher, R. A., *Proc. Camb. Phil. Soc.*, *26*, 528, (1930).

[9] Fisher, R. A., *Proc. Roy. Soc.*, *A 139*, 343, (1933).

replace the unknown $f(x)$. Obviously, $\overline{P}(x_k < x < x_{k+1}) = \dfrac{1}{n+1}$. A discussion of the special case, $n = 2$, has been prominent recently in a controversy between Jeffrey[10] and Fisher[9,11] and in an article by Bartlett.[12]

Now, in (3) for $p = p'$, and $p'' = 1$; we may write[13]

$$(7) \quad \overline{P}(p < p_k) = \sum_{\alpha=0}^{k-1} \binom{n}{\alpha} \cdot p^{\alpha} \cdot q^{n-\alpha} \equiv I_q(n-k+1, k) \equiv \frac{B_q(n-k+1, k)}{B_1(n-k+1, k)},$$

where $q = 1 - p$, and the incomplete $B$ and $I$ functions are those of K. Pearson[13] and Müller.[14] Now, let $M$ be the unknown median of the infinite population, $U$. Then, by definition of $p_k$, if and only if $x_k > M$, then $p_k > \frac{1}{2}$. Therefore,

$$(8) \quad \overline{P}(M < x_k) = \overline{P}(0.5 < p_k) = \left(\frac{1}{2}\right)^n \cdot \sum_{\alpha=0}^{k-1} \binom{n}{\alpha} \equiv I_{0.5}(n-k+1, k).$$

Obviously, $\overline{P}(x_k < M < x_{k+1}) = \left(\frac{1}{2}\right)^n \cdot \binom{n}{k}$, and the expectation that $M$ lie between the $k$-th observations from each end of the set, $S$, is given by

$$(9) \quad \overline{P}(x_k < M < x_{n-k+1}) = 1 - 2 \cdot I_{0.5}(n-k+1, k), \text{ for } 2k < n + 1.$$

Obviously, this confidence range is *central.*

## Section 2

Now, consider another infinite population, $U'$. In similar manner we may develop expressions for confidence ranges and distribution expectations. Let $x'$ be the variate, and consider a sample, $S' = \{x'_m\}$, of $n'$ observations drawn *without replacements* from $A$ according to $U'$ but after the sample, $S$, of $U$; i.e., so that no two of these sample values in $S'$ are equal, nor any of them equal to a value in $S$. Furthermore, let $m$ be the order of ascending magnitude of $x'$ values in $S'$; and $p'_m \equiv P(x' < x'_m)$ for $x'$ drawn at random from $U'$, and let $M'$ be the unknown median of $U'$. Then, by replacement of $x$, $n$, $p_k$, $k$, and $M$ by $x'$, $n'$, $p'_m$, $m$, and $M'$, respectively, in relations already developed for $U$ and $S$, we obtain corresponding expressions for $U'$ and $S'$; e.g.,

$$(10) \quad \overline{P}(x'_m < x' < x'_{m+1}) = \frac{1}{n'+1}.$$

[10] Jeffreys, H., *Proc. Roy. Soc., A 138*, 48, (1932); *A 140*, 523, (1933); *A 146*, 9, (1934); *Proc. Camb. Phil. Soc., 29*, 83, (1933).

[11] Fisher, R. A., *Proc. Roy. Soc., A 146*, 1, (1934).

[12] Bartlett, M. S., *Proc. Roy. Soc., A 141*, 518, (1933).

[13] Pearson, K., *Biometrika, 16*, 202, (1924).

[14] Müller, J. H., *Biometrika, 22*, 284, (1930–31).

Now, let the index values, $k_m$, be defined as the number of values of $\{x_k\}$ that are less than $x'_m$, $m = 1, \cdots, n'$. Then, for all realized cases,

$$(11) \qquad x_{k_m} < x'_m < x_{k_m+1}, \qquad m = 1, \cdots, n',$$

for the extreme members of (11) in $S$. Then, for $x$ and $x'$ drawn at random from $U$ and $U'$, respectively, we may write

$$(12) \qquad 0 < (n+1)(n'+1)\cdot\overline{P}(x < x') - \sum_{m=1}^{n'} k_m < n + n' + 1,$$

provided that the expectations for $U$ and $U'$ may be treated as independent. Similarly, for $\overline{P}(M < M')$ we have the relations,

$$(13) \qquad \begin{aligned} \sum_{m=1}^{n'} \binom{n'}{m}\cdot I_{0.5}(n - k_m + 1, k_m) < 2^{n'}z\cdot\overline{P}(M < M') < 1 \\ + \sum_{m=1}^{n'} \binom{n'}{m-1}\cdot I_{0.5}(n - k_m, k_m + 1). \end{aligned}$$

Of course, $I_{0.5}(n + 1, 0) \equiv 0$, and $I_{0.5}(0, n + 1) \equiv 1$. It may be verified readily that the inequality relations of (12) and (13) provide *best* upper and lower bounds for $\overline{P}(x < x')$ and $\overline{P}(M < M')$ under the circumstances given.

Obviously, any increasing function, $\phi(y)$, for $y$ in $A$, may be used throughout the arguments, with $\phi(y)$ replacing $y = x, x_k, M, x', x'_m, M'$, respectively.

## Section 3

Consider, now, the case of a finite population, $U_N$, of real numbers $\{x^{(i)}\}$, $x^{(i)} < x^{(j)}$ for $i < j$, $i = 1, \cdots, N$. Assume that $N$ is known, and that a sample, $S$, of $n$ values has been drawn at random from $U_N$ without replacements. Let the sample values be $\{x_k\}$, $k = 1, \cdots, n$; and $k$ be an arbitrarily determined index. As before, we might consider $k$ the order of draft, temporarily, but the same analysis may be made if we let $k$ be the order of ascending magnitude in the sample, $S$, and disregard its value in connection with *à priori* estimates of draft probability. Each $x_k = x^{(u_k)}$ for some unknown $u_k = 1, \cdots, N$; and, *à priori* (i.e., with no knowledge as to order of magnitude of other values in the sample), any two of these values are equally likely. Obviously, this is so if $x_k$ is the first value drawn from $U_N$, and the rest of the sample may be regarded as a random draft without replacements of $n - 1$ elements from $[U_N - x_k]$. Let $r$ be the number of these sample values less than $x_k$, and $s = n - 1 - r$. Then the probability of drawing such a sample after the given $x_k$, under the conditions given, is $\dfrac{\binom{u_k - 1}{r}\binom{N - u_k}{s}}{\binom{N-1}{n-1}}$, where $u_k - 1$ is the unknown number of values in $U_N$ that are less than $x_k$. To estimate the expectation, $\overline{P}(R = u_k - 1)$, that there are just a given number, $R$, of values in $U_N$ less than $x_k$; we encounter the same situation considered by K. Pearson in a paper[15] subsequent to those applied to the infinite universe; and, by a simple conversion in notation, we have

$$(14) \qquad \overline{P}(R = u_k - 1) = \frac{\binom{R}{r} \cdot \binom{N - 1 - R}{s}}{\binom{N}{n}}.$$

In previous communications[16,17] I have defined a function,

$$(15) \qquad \psi(r, s, r', s') \equiv \frac{\sum_{\alpha=0}^{r'} \binom{r + r' - \alpha}{r} \cdot \binom{s + s' + 1 + \alpha}{s}}{\binom{r + s + r' + s' + 2}{r + s + 1}},$$

for any four rational integers $r, s, r', s' \geqq 0$; and shown that Pearsons further result, equivalent here to evaluation of $\overline{P}(u_k \leqq R + 1)$ for a given $R$, may be expressed by means of this $\psi$-function. Thus, we have

$$(16) \qquad \overline{P}(u_k \leqq R + 1) = \psi(r, s, R - r, N - R - s - 2).$$

It was demonstrated also[16,17] that

$$(17) \qquad \psi(r, s, r', s') \equiv \psi(r, r', s, s') \equiv \psi(s', r', s, r) \equiv 1 - \psi(s, r, s', r')$$

with extension of the definition to include $\psi(r, s, -1, s') \equiv 0$, and that

$$(18) \qquad \psi(r, s, r', s') \equiv \frac{\sum_{\alpha=0}^{\alpha' \leqq s, r'} \binom{r + r' + 1}{r + 1 + \alpha} \cdot \binom{s + s' + 1}{s - \alpha}}{\binom{r + s + r' + s' + 2}{r + s + 1}}.$$

As in the case of the infinite population, here also it is obvious that the order of draft of $x_k$ is of no consequence in the analysis; and again we will let $k = r + 1$, whence $s = n - k$, and we may make these substitutions in (14) and (16). Then, we may write

$$(19) \qquad \overline{P}(u_k \leqq R) = \psi(k - 1, n - k, R - k, k + N - R - n - 1);$$

[15] Pearson, K., *Biometrika, 20 A*, 149, (1928).

[16] Thompson, W. R., *Biometrika, 25*, 285, (1933).

[17] Thompson, W. R., *American Journal of Mathematics*, *57*, 450, (1935).

and, obviously, $\overline{P}(u_{n-k+1} \geqq N - R + 1) \equiv \overline{P}(u_k \leqq R)$. Hence, if we let $M$ be the unknown median of $U_N$; and $m \equiv \frac{N-a}{2}$, where $a = 0, 1$, and $N - a$ is even; then, as $u_i$ is an integer,

$$\overline{P}(x_k \leqq M \leqq x_{n-k+1}) \equiv \overline{P}\left(u_k \leqq \frac{N}{2} \leqq u_{n-k+1}\right) \tag{20}$$

$$\equiv 1 - 2\cdot\psi(k-1, n-k, m-k, k+N-m-n-1),$$

which is the expectation that the median of $U_N$ lie within the closed interval, $(x_k, x_{n-k+1})$, for $2k \leqq n + 1$. This gives the confidence range, analogous to that for the infinite universe. It may be noted that

$$\overline{P}(u_k \leqq R < u_{k+1}) = \overline{P}(u_k \leqq R) - \overline{P}(u_{k+1} \leqq R)$$

$$= \psi(r, s, r', s') - \psi(r+1, s-1, r'-1, s'+1)$$

where $r = k - 1$, $s = n - k$, $r' = R - k$, and $s' = k + N - R - n - 1$. Hence, (18) gives

$$\overline{P}(u_k \leqq R < u_{k+1}) \equiv \frac{\binom{R}{k}\cdot\binom{N-R}{n-k}}{\binom{N}{n}}. \tag{21}$$

The approach by way of Pearson's problem again makes it easy to evaluate the expected mean $p_k$ and variance as in the case of the infinite population, where $p_k = P(x < x_k)$ for $x$ drawn at random from $U_N$. Of course, $p_k = \frac{u_k - 1}{N}$, but $u_k$ is unknown. From Pearson's result,[15] however, we obtain

$$\bar{p}_k = \frac{k(N+1) - n - 1}{N(n+1)} = \frac{k}{n+1}\left(1 - \frac{n}{N}\right) + \frac{k-1}{N}, \tag{22}$$

and the expected variance of $p_k$,

$$\overline{\sigma^2_{p_k}} = E((p_k - \bar{p}_k)^2) = \frac{k(n-k+1)(N+1)(N-n)}{(n+1)^2\cdot(n+2)\cdot N^2}. \tag{23}$$

Yale University.

# THE SAMPLING DISTRIBUTION OF THE COEFFICIENT OF VARIATION

By Walter A. Hendricks with the assistance of Kate W. Robey

National Agricultural Research Center, Beltsville, Maryland

The coefficient of variation does not appear to be of very great interest to statisticians in general. However, its use in biometry is sufficiently extensive for some knowledge of its sampling distribution to be desirable. The present paper is an attempt to satisfy this need.

For the purposes of the following discussion, the coefficient of variation may be defined as the ratio of the standard deviation of a number of measurements to the arithmetic mean:

$$v = \frac{s}{\bar{x}} \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (1)$$

As is well known, the probability that the mean of a sample of $n$ measurements, taken at random from a normal universe, lies between $\bar{x}$ and $\bar{x} + d\bar{x}$ and that the standard deviation of the measurements in the same sample lies between $s$ and $s + ds$ is given by the relation:

$$dF_{\bar{x},s} = \frac{n^{\frac{1}{2}n}}{2^{\frac{1}{2}n-1}\,\pi^{\frac{1}{2}}\,\Gamma\left(\frac{n-1}{2}\right)\sigma^n}\, e^{-\frac{n}{2\sigma^2}[(\bar{x}-m)^2+s^2]}\, s^{n-2}\, d\bar{x}\, ds \ldots\ldots (2)$$

If equation (2) is expressed in terms of polar coördinates by means of the transformation: $\bar{x} = \rho \cos \theta$; $s = \rho \sin \theta$, it becomes a distribution function of $\rho$ and $\theta$ in which $\theta = \text{arc tan } v$:

$$dF_{\rho,\theta} = \frac{n^{\frac{1}{2}n}}{2^{\frac{1}{2}n-1}\,\pi^{\frac{1}{2}}\,\Gamma\left(\frac{n-1}{2}\right)\sigma^n}\, e^{-\frac{n}{2\sigma^2}(\rho^2-2m\rho\cos\theta+m^2)}\, \rho^{n-1} \sin^{n-2}\theta\, d\rho\, d\theta \ldots (3)$$

In equation (3), $\rho$ may vary from 0 to $\infty$ and $\theta$ may vary from 0 to $\pi$. To find the distribution function of $\theta$, all that is necessary is to write:

$$dF_\theta = k\left[\int_0^\infty e^{-(a\rho-b)^2}\,\rho^{n-1}\,d\rho\right] d\theta \ldots\ldots\ldots\ldots (4)$$

in which,

$$k = \frac{n^{\frac{1}{2}n}}{2^{\frac{1}{2}n-1}\,\pi^{\frac{1}{2}}\,\Gamma\left(\frac{n-1}{2}\right)\sigma^n}\,e^{-\frac{n}{2\sigma^2}m^2\sin^2\theta}\sin^{n-2}\theta,$$

$$a = \frac{n^{\frac{1}{2}}}{2^{\frac{1}{2}}\,\sigma}, \quad \text{and} \quad b = \frac{n^{\frac{1}{2}}}{2^{\frac{1}{2}}\,\sigma}\,m\cos\theta,$$

and to perform the indicated integration.

To evaluate the integral inside the brackets in equation (4), we may write:

$$\int_0^\infty e^{-(a\rho-b)^2}\rho^{n-1}\,d\rho = \frac{1}{a^n}\int_{-b}^\infty e^{-u^2}\sum_{i=0}^{n-1}\frac{(n-1)!}{(n-1-i)!\,i!}\,u^{n-1-i}\,b^i\,du\ldots.(5)$$

Consider the integral, $\int_{-b}^\infty e^{-u^2}\,u^{n-1-i}\,du$. If $b$ is sufficiently large, as is the case when the parameters of equation (2) are of such magnitude that practically the entire volume under the frequency surface lies to the right of the $s$ axis, that is to say, if negative and small positive values of $\bar{x}$ occur so infrequently that their effects may be neglected, the lower limit, $-b$, of this integral may be replaced by $-\infty$ without introducing any appreciable error. The value of the integral, $\int_{-\infty}^\infty e^{-u^2}\,u^{n-1-i}\,du$, is zero when $n-1-i$ is odd and $\Gamma\left(\frac{n-i}{2}\right)$ when $n-1-i$ is even, zero being counted as an even number.

Subject to the above condition that $b$ be sufficiently large, we may, therefore, write equation (5) in the form:

$$\int_0^\infty e^{-(a\rho-b)^2}\rho^{n-1}\,d\rho = \frac{1}{a^n}\sideset{}{'}\sum_{i=0}^{n-1}\frac{(n-1)!}{(n-1-i)!\,i!}\,\Gamma\left(\frac{n-i}{2}\right)b^i\ldots\ldots(6)$$

in which the symbol, $\sum'$, indicates that the only terms entering into the summation are those in which $n-1-i$ is an even number.

Substituting this expression for the integral inside the brackets in equation (4), replacing $k$, $a$, and $b$ by the quantities which they represent, and writing $V$ in place of the ratio, $\frac{\sigma}{m}$, we obtain the following distribution function of $\theta$:

$$dF_\theta = \frac{2}{\pi^{\frac{1}{2}}\,\Gamma\left(\frac{n-1}{2}\right)}\,e^{-\frac{n}{2V^2}\sin^2\theta}\sin^{n-2}\theta\sideset{}{'}\sum_{i=0}^{n-1}\frac{(n-1)!\,\Gamma\left(\frac{n-i}{2}\right)}{(n-1-i)!\,i!}\,\frac{n^{\frac{1}{2}i}}{2^{\frac{1}{2}i}\,V^i}\cos^i\theta\,d\theta.(7)$$

Equation (7) may be written in terms of $v$, if desired, by making the substitution, $\theta = \text{arc}\tan v$:

$$dF_v = \frac{2}{\pi^{\frac{1}{2}}\,\Gamma\left(\frac{n-1}{2}\right)}\, e^{-\frac{n}{2V^2}\frac{v^2}{1+v^2}}\, \frac{v^{n-2}}{(1+v^2)^{\frac{1}{2}n}}$$

$$\sum_{i=0}^{n-1}{}' \frac{(n-1)!\,\Gamma\left(\frac{n-i}{2}\right)}{(n-1-i)!\,i!}\, \frac{n^{\frac{1}{2}i}}{2^{\frac{1}{2}i}\,V^i}\, \frac{1}{(1+v^2)^{\frac{1}{2}i}}\, dv \,.\,.\; (8)$$

It must be emphasized that equation (8) has been derived on the hypothesis that negative and small positive values of $\bar{x}$ occur so infrequently that they may be neglected. However, since this condition is satisfied in the vast majority of practical problems in which the coefficient of variation is likely to be used, the limitation is not of much practical importance.

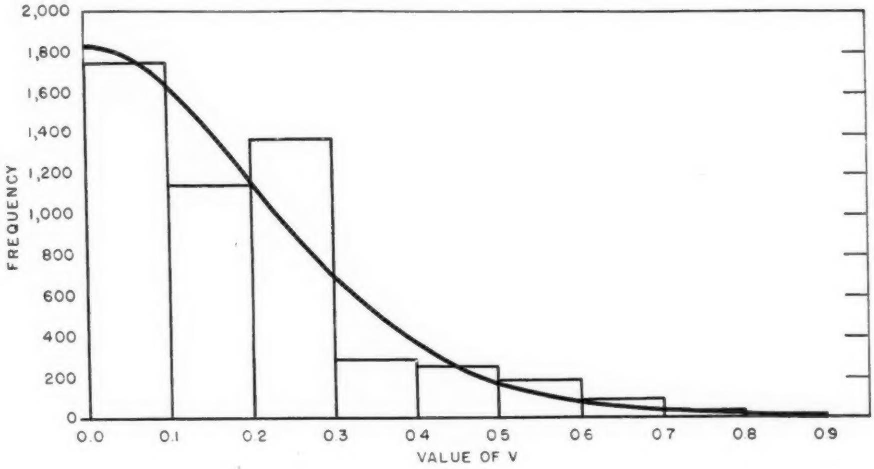

FIG. 1. OBSERVED AND THEORETICAL DISTRIBUTIONS OF VALUES OF $v$ FOR 512 SAMPLES OF NUMBERS OF HEADS APPEARING IN TWO SUCCESSIVE TOSSES OF TEN COINS

As a test of the validity of equation (8), the authors calculated 512 coefficients of variation of the numbers of heads appearing in two successive tosses of ten coins. The coins were tossed 1024 times, thus yielding 512 samples, each consisting of two observations. For these data we have $m = 5$, $\sigma = 1.581$, and $V = 0.3162$.

For the case, $n = 2$, equation (8) reduces to:

$$dF_v = \frac{2}{\pi^{\frac{1}{2}}V}\, e^{-\frac{1}{V^2}\frac{v^2}{1+v^2}}\, \frac{dv}{(1+v^2)^{\frac{3}{2}}} \ldots\ldots\ldots\ldots\ldots\ldots (9)$$

Figure 1 shows the distribution of the 512 values of $v$ obtained from the coin tossing experiment, together with the theoretical distribution given by equation (9).

An inspection of Figure 1 indicates that the agreement between the observed and theoretical frequencies is fairly good. An application of the familiar chi test for goodness of fit showed the agreement to be rather poor. According to this test, the degree of discrepancy between theory and observation could have arisen by chance less than once in a hundred trials. However, the discrepancies may be partly due to the fact that data distributed in a discrete fashion were treated by methods appropriate to the analysis of data distributed according to a continuous frequency curve.

As another test of the validity of equation (8), the authors calculated 149 coefficients of variation of "days to maturity," which is the length of time elapsing between the date of hatch of a chicken and the time egg production commences, for samples of two observations made upon Rhode Island Red pullets. Figure 2 shows the observed distribution of the 149 coefficients of variation, together with the theoretical distribution given by equation (9).

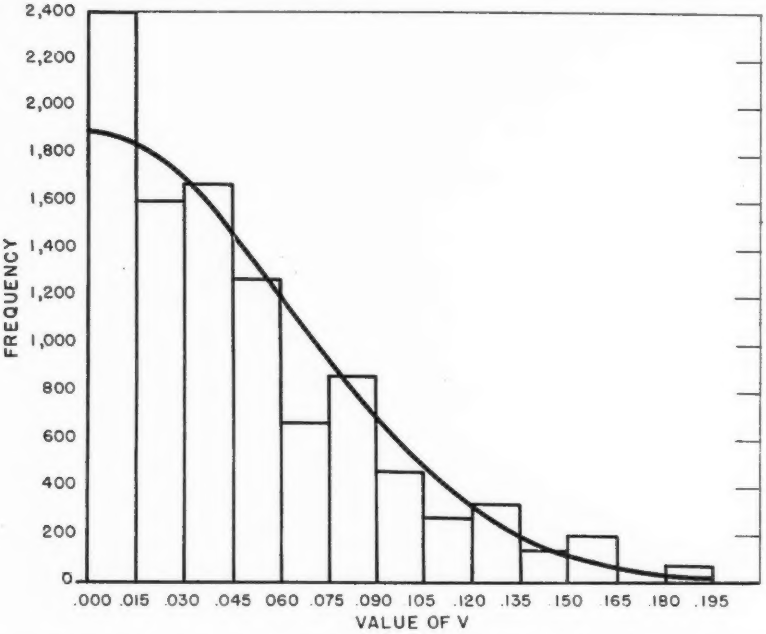

FIG. 2. OBSERVED AND THEORETICAL DISTRIBUTIONS OF VALUES OF $v$ FOR 149 SAMPLES OF "DAYS TO MATURITY" IN RHODE ISLAND RED PULLETS FOR SAMPLES OF TWO OBSERVATIONS

In applying equation (9) to these data, the parameter, V, had to be evaluated from the data. The best estimates of the values of $m$, $\sigma$, and $V$ which could be obtained from the 298 measurements of "days to maturity" are $m = 210.477$, $\sigma = 18.6991$, $V = 0.0888415$. The theoretical distribution shown in Figure 2 is based on this value of $V$.

The agreement between theory and observation shown by Figure 2 is very good. In this case, the chi test showed that the degree of discrepancy encountered could have arisen by chance about six times in ten trials.

# SOME NOTES ON EXPONENTIAL ANALYSIS

By H. R. Grummann

Assistant Professor, Department of Applied Mathematics, Washington University

M. E. J. Geuhry de Bray in his charming little book "Exponentials made Easy"[1] tells how to determine the constants in the equation,

$$\text{(I)} \qquad y = A_1\epsilon^{a_1x} + A_2\epsilon^{a_2x}$$

so that the curve will pass through four points, with equidistant ordinates on an empirical curve. If (Fig. 1) $y_0$, $y_1$, $y_2$, and $y_3$ are the equidistant ordinates and $\delta$ is their common separation, $y_0$ being the $y$ intercept of the curve, de Bray's formulas are:

$$\text{(II)} \qquad a_1 = \frac{\log z_1}{\delta}, \qquad a_2 = \frac{\log z_2}{\delta}$$

where $z_1$ and $z_2$ are the roots of the quadratic equation

$$\text{(III)} \qquad \begin{vmatrix} z^2 & z & 1 \\ y_3 & y_2 & y_1 \\ y_2 & y_1 & y_0 \end{vmatrix} = 0.$$

The coefficients $A_1$ and $A_2$ of the two exponential terms are obtained by solving the two simultaneous equations

$$\text{(IV)} \qquad \begin{aligned} A_1 + A_2 &= y_0 \\ A_1z_1 + A_2z_2 &= y_1 \end{aligned}$$

In attempting to find suitable empirical equations for some "river rating curves"—graphs of discharge versus stage—the writer tried to make use of de Bray's procedure. The original intention was to use the above method to determine the constants, and then to correct these constants by the use of Least Squares, as done by J. W. T. Walsh[2] in an application of the method to a problem in radioactivity. It often happens that a series of plotted observations suggest a simple exponential function, but that when the observations are replotted on semi-logarithmic paper a straight line is not obtained. Often, as in the case of a good many river rating curves, the result may be described

[1] Macmillan & Co. Ltd., St. Martin's St., London W. C. 2.

[2] Proceedings Phys. Soc. London XXXII. This reference is given by de Bray in his book, "Exponentials made Easy."

as "almost straight." At first blush it might seem that in all such cases it ought to be possible to fit a curve with equation I to the data by de Bray's Method. By an easy generalization of the above formulas, the constants in an equation with three or four exponential terms could be determined if two terms were not enough to secure a good fit.

It was soon found, however, that innocent looking monotonic curves without points of inflection plotted from data that gave an "almost straight" line on semi-logarithmic paper quite often led to a quadratic equation, (equation III) whose roots were not both positive numbers.

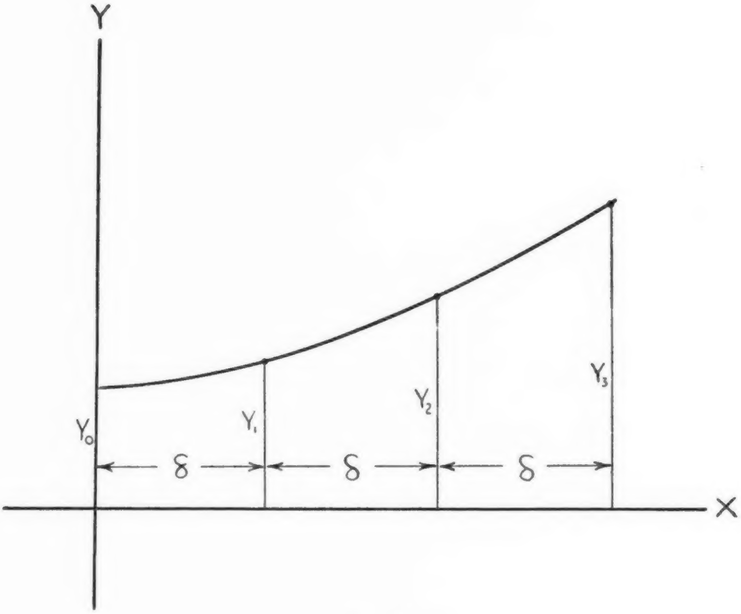

FIG. 1

If $z_1$ and $z_2$, the roots of III, are complex conjugates, it may be seen from IV that $A_1$ and $A_2$ will be complex conjugates. Also, $a_1$ and $a_2$ will be conjugate complex numbers and may be calculated as follows:

Let $z_1 = r\epsilon^{\iota\theta}$ and $z_2 = r\epsilon^{-\iota\theta}$.
then from equation II,

$$r\epsilon^{\iota\theta} = \epsilon^{a_1\delta},$$

$$r\epsilon^{-\iota\theta} = \epsilon^{a_2\delta}$$

whence, by division to eliminate $r$ we have

$$\epsilon^{2\iota\theta} = \epsilon^{\delta(a_1-a_2)}, \text{ or}$$

$$\text{(Va)} \qquad \frac{2i\theta}{\delta} = a_1 - a_2.$$

Also, by multiplication to eliminate $\theta$,

$$r^2 = \epsilon^{\delta(a_1+a_2)}, \text{ or}$$

$$\text{(Vb)} \qquad \frac{2 \log r}{\delta} = a_1 + a_2 .$$

The sum and difference of the two $a$'s being obtained by these expressions, one may solve for $a_1$ and $a_2$.

Let
$$a_1 = \lambda + \iota\mu \qquad A_1 = \alpha + \iota\beta$$
$$a_2 = \lambda - \iota\mu \qquad A_2 = \alpha - \iota\beta$$

Then equation I becomes

$$y = (\alpha + \iota\beta)\epsilon^{(\lambda+\iota\mu)x} + (\alpha - \iota\beta)\epsilon^{(\lambda-\iota\mu)x},$$
$$y = 2\epsilon^{\lambda x}[\alpha \cos \mu x - \beta \sin \mu x], \text{ or}$$
$$\text{(VI)} \qquad y = 2\epsilon^{\lambda x} R \cos (\mu x + c)$$

where $R = \sqrt{\alpha^2 + \beta^2}$ and $\tan c = \dfrac{\beta}{\alpha}$.

If one of the roots of III is negative, the de Bray formulas II and IV will still give an expression for equation I which formally reproduces $y_0$, $y_1$, $y_2$, and $y_3$ when 0, $\delta$, $2\delta$, and $3\delta$, are substituted for $x$ respectively, but which is useless for interpolating and of no value as a solution of the curve fitting problem. Suppose, for example, that $z_1$ is positive and $z_2$ is negative. Then

$$z_2 = (-1) \mid z_2 \mid \qquad \text{and}$$

$$\log z_2 = \log (-1) + \log \mid z_2 \mid .$$

Equation I then becomes

$$y = A_1\epsilon^{a_1x} + (-1)^{\frac{x}{\delta}} A_2\epsilon^{\frac{x \log \mid z_2 \mid}{\delta}},$$

the factor $(-1)^{\frac{x}{\delta}}$ being real only when $x$ is an integral multiple of $\delta$. If the $(-1)$ is written $e^{\iota\pi}$, we have

$$y = A_1\epsilon^{a_1x} + \epsilon^{\frac{\pi \iota x}{\delta}} A_2\epsilon^{\frac{x \log \mid z_2 \mid}{\delta}}, \text{ or}$$

$$y = A_1\epsilon^{a_1x} + A_2\epsilon^{\frac{x \log \mid z_2 \mid}{\delta}}\left[\cos \frac{\pi x}{\delta} + \iota \sin \frac{\pi x}{\delta}\right].$$

Neither the real nor the imaginary part would be a graduation function for a monotonic curve as each has a half period of $\delta$.

The expression for I is similar, and of no greater practical value, if both of the roots of III are negative.

Without loss of generality we may let $y_0 = 1$, $r_1 = \frac{y_1}{y_0}\ r_2 = \frac{y_2}{y_1}\ r_3 = \frac{y_3}{y_2}$. Then the quadratic III becomes

$$\begin{vmatrix} z^2 & z & 1 \\ r_2 r_3 & r_2 & 1 \\ r_1 r_2 & r_1 & 1 \end{vmatrix} = 0, \qquad \text{or,}$$

written in the form

$$z^2 + pz + q = 0, \quad \text{i.e.,}$$

$$\text{(IIIa)} \qquad z^2 + \frac{r_2(r_1 - r_3)}{(r_2 - r_1)} z + \frac{r_1 r_2 (r_3 - r_2)}{(r_2 - r_1)} = 0.$$

Hence the roots of this quadratic are real and unequal if $D > 6$, equal if $D = 6$, and complex if $D < 6$, where

$$D = \left[\frac{r_3}{r_1} - 3\frac{r_1}{r_3}\right] + 4\left[\frac{r_1}{r_2} + \frac{r_2}{r_3}\right]$$

From the point of view of the computer, however, it is about as much work to calculate $D$ as to solve the quadratic equation.

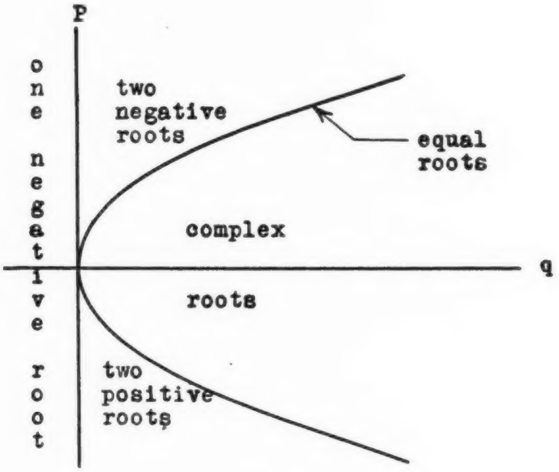

FIG. 2

Reverting to equation IIIa; suppose the numbers $q$ and $p$ are plotted as the coordinates of a point $(q, p)$ as in Fig. 2. Then the parabola $p^2 = 4q$ is, so to speak, a locus of equal roots. The remainder of the figure requires no explanation.

Suppose that all the $r$'s are positive, as they would be in the case of a simple monotonic curve which one proposed subjecting to an exponential analysis.

If $q < 0$, the quadratic will have one negative root. Now

$$q = \frac{r_1 r_2(r_3 - r_2)}{(r_2 - r_1)} \quad \text{and hence}$$

for $q < 0$, if $r_2 > r_1$, then $r_3 < r_2$ and consequently $r_3 < r_2 > r_1$ and if $r_2 < r_1$, then $r_3 > r_2$, or $r_1 > r_2 < r_3$. Also, provided $p_2 > 4q$, a positive $p$ and a positive $q$ will give two negative roots. But

$$p = \frac{r_2(r_1 - r_3)}{(r_2 - r_1)},$$

and $p$ and $q$ can not both be positive when all the $r$'s are positive as this implies either that $r_2 > r_1$, $r_1 > r_3$ and $r_3 > r_2$, a contradiction, or else that $r_2 < r_1$, $r_1 < r_3$ and $r_3 < r_2$, also a contradiction. Hence if both roots are negative, the $r$'s can not be all positive. The case of two negative roots will not arise in trying to fit equation I to a monotonic curve, since if all the $r$'s are positive both $p$ and $q$ can not be positive.

For all $r$'s positive, provided $p^2 > 4q$, a positive $q$ and a negative $p$ will give two positive roots. But

$$q = \frac{r_1 r_2(r_3 - r_2)}{(r_2 - r_1)} > 0,$$

and

$$-p = \frac{r_2(r_3 - r_1)}{(r_2 - r_1)} > 0$$

means that $r_3 > r_2 > r_1$ or $r_3 < r_2 < r_1$.

To sum up: If all the $r$'s are positive, de Bray's method of exponential analysis is possible (a) when $D < 6$ and the roots of III are complex; (b) when $D > 6$ and $r_1 > r_2 > r_3$ or when $r_1 < r_2 < r_3$.

Figure 3 gives a picture of the second condition (b) of the preceding paragraph. Suppose an exponential curve is passed through the first *two* points on the empirical curve with ordinates $y_0$ and $y_1$. Its equation will be:

$$y = y_0\left(\frac{y_1}{y_0}\right)^{\frac{x}{\delta}} = y_0 r_1^{\frac{x}{\delta}}.$$

Suppose also that $y_2$ is less than the ordinate to this curve when $x = 2\delta$. Now pass an exponential curve through $y_1$ and $y_2$ using a new axis of ordinates coinciding with $y_1$. Its equation is

$$y = y_1\left(\frac{y_2}{y_1}\right)^{\frac{x}{\delta}} = y_1 r_2^{\frac{x}{\delta}},$$

or referred to the original axis:

$$y = y_1 r_2^{\frac{x-\delta}{\delta}}.$$

Now if the graduation is possible without using trigonometric functions, $y_3$ must be less than the ordinate of this second curve when $x = 3\delta$.

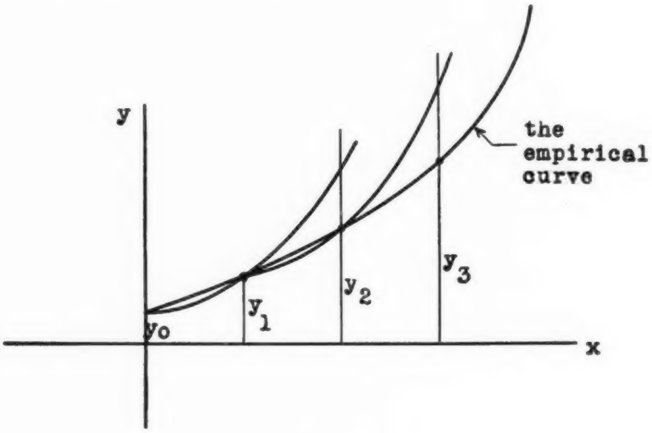

FIG. 3

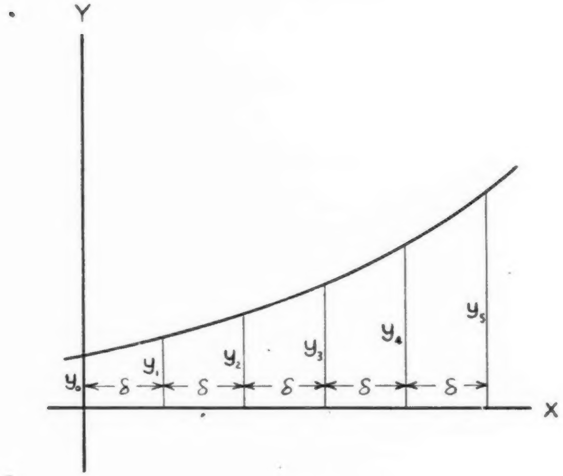

FIG. 4

It is natural to inquire if the state of affairs is not similar to this, for the cases of fitting curves with equations similar to I but having three or four exponential terms on the right hand side instead of only two. If three terms are used (see Fig. 4) to find constants in

(Ia) $$y = A_1\epsilon^{a_1x} + A_2\epsilon^{a_2x} + A_3\epsilon^{a_3x}$$

it is first necessary to find the roots of the cubic

$$\text{(IIIa)} \qquad f(x) = \begin{vmatrix} z^3 & z^2 & z & 1 \\ y_5 & y_4 & y_3 & y_2 \\ y_4 & y_3 & y_2 & y_1 \\ y_3 & y_2 & y_1 & y_0 \end{vmatrix}.$$

Now, $f(x)$ will have no negative roots if $f(-x)$ has no changes of sign. But writing the conditions that the cofactors of the elements of the first row in the above determinant have the same signs, and assuming that all the $y$'s are positive, one does not get a series of conditions analogous to $r_3 > r_2 > r_1$ or $r_3 < r_2 < r_1$.

In the following, formulas will be derived for finding the constants in equation Ia after the roots of IIIa have been determined. Also formulas will be obtained for finding the constants in

$$\text{(Ib)} \qquad y = A_1\,\epsilon^{a_1x} + A_2\,\epsilon^{a_2x} + A_3\,\epsilon^{a_3x} + A_4\,\epsilon^{a_4x}$$

after the roots of

$$\text{(IIIb)} \qquad \begin{vmatrix} z^4 & z^3 & z^2 & z & 1 \\ y_7 & y_6 & y_5 & y_4 & y_3 \\ y_6 & y_5 & y_4 & y_3 & y_2 \\ y_5 & y_4 & y_3 & y_2 & y_1 \\ y_4 & y_3 & y_2 & y_1 & y_0 \end{vmatrix} = 0$$

have been found. Both sets of formulas have been tested by an "exponential analysis" of the same body of data, viz., the very accurate recent determinations by the U. S. Bureau of Standards of the saturation pressure of water vapor above 100C.[3]

For the case of three exponential terms in the graduation function, the $a$'s are found by formulas like II or V, after the roots of the cubic are found. If $z_1$, $z_2$, $z_3$ are the roots, the $A$'s are obtained by solving the simultaneous equations

$$\text{(IVa)} \qquad \begin{aligned} A_1 + A_2 \quad + A_3 \quad &= y_0 \\ A_1z_1 \;+ A_2z_2 \;+ A_3z_3 \;&= y_1 \\ A_1z_1{}^2 + A_2z_2^2 + A_3z_3^2 &= y_2 \end{aligned}$$

[3] Osborne, Stimson, Fiock, and Ginnings: The Pressure of Saturated Water Vapor in the Range 100° to 374°C. Bureau Standards Journal of Research, Vol. 10, Febr. 1933, page 178.

This presents no new difficulty unless two of the roots are conjugate complex numbers. In this event, if we let $z_1 =$ the real positive root, $z_2 = r\,\epsilon^{\iota\theta}$, and $z_3 = r\,\epsilon^{-\iota\theta}$ the determinant $D$ of the equations IVa may be written

$$D = \begin{vmatrix} 1 & 1 & 1 \\ z_1 & r\epsilon^{\iota\theta} & r\epsilon^{-\iota\theta} \\ z_1^2 & r^2\epsilon^{2\iota\theta} & r^2\epsilon^{-2\iota\theta} \end{vmatrix}$$

or, expanded in terms of the elements of the first column and their minors,

$$D = 2i[z_1 r^2 \sin 2\theta - (r^3 + z_1^2 r)\sin \theta],$$

a pure imaginary. Similarly,

$$A_1\,D = 2i[r^2 y_1 \sin 2\theta - (y_0 r^3 + y_2 r)\sin \theta],$$

also a pure imaginary, so that $A_1$ is real. Having calculated $A_1$, it is substituted in the first two of equations IVa, which are then solved for $A_2$ and $A_3$. $a_2$ and $a_3$ are then determined by formulas Va and Vb, replacing the subscripts 1 and 2 in those formulas, by the subscripts 2 and 3 respectively. Finally the two exponential terms corresponding to the complex roots of the cubic are combined into a single trigonometric term as in equation VI.

The necessary formulas for the case of four exponential terms in the graduation function will be discussed briefly. The equations

$$\text{(IVb)} \qquad \begin{aligned} A_1 &+ A_2 &+ A_3 &+ A_4 &= y_0 \\ A_1 z_1 &+ A_2 z_2 &+ A_3 z_3 &+ A_4 z_4 &= y_1 \\ A_1 z_1^2 &+ A_2 z_2^2 &+ A_3 z_3^2 &+ A_4 z_4^2 &= y_2 \\ A_1 z_1^3 &+ A_2 z_2^3 &+ A_3 z_3^3 &+ A_4 z_4^3 &= y_3 \end{aligned}$$

have to be solved for the $A$'s. The $z$'s are the roots of IIIb. Two cases will be considered: First case: $z_1$ and $z_2$ are complex conjugates and $z_3$ and $z_4$ are complex conjugates. Second case: $z_1$ and $z_2$ are complex conjugates and $z_3$ and $z_4$ are real and positive. In either event $A_1$ and $A_2$ are complex conjugates, as will be proved below. Formulas for $A_1$ are given for both cases. Then $A_2$ is known since it is the conjugate of $A_1$. Having found $A_1$ and $A_2$, let

$$\begin{aligned} c_0 &= y_0 - (A_1 + A_2) \\ c_1 &= y_1 - (A_1 z_1 + A_2 z_2) \end{aligned}$$

Both $c_0$ and $c_1$ are then real. To get $A_3$ and $A_4$ solve the equations:

$$\begin{aligned} A_3 \;+ A_4 &= c_0 \\ A_3 z_3 + A_4 z_4 &= c_1 \end{aligned}$$

A pair of exponential terms with conjugate complex coefficients will then be expressed as a single real trigonometric term as in VI.

The determinant of equations IVb may be written

$$\text{(VII)}\qquad D = (z_1 - z_2)(z_1 - z_3)(z_1 - z_4)(z_2 - z_3)(z_2 - z_4)(z_3 - z_4).$$

First case: Let $z_1 = a + \iota b$, $z_2 = a - \iota b$, $z_3 = \alpha + \iota\beta$, $z_4 = \alpha - \iota\beta$. Then $D$ may be written

$$\text{(VIIa)}\qquad D = -4\beta b[(a - \alpha)^2 + (b - \beta)^2]\,[(a - \alpha)^2 + (b + \beta)^2],$$

which is real. Now

$$A_1D + A_2D = \begin{vmatrix} y_0 & 1 & 1 & 1 \\ y_1 & z_2 & z_3 & z_4 \\ y_2 & z_2^2 & z_3^2 & z_4^2 \\ y_3 & z_2^3 & z_3^3 & z_4^3 \end{vmatrix} + \begin{vmatrix} 1 & y_0 & 1 & 1 \\ z_1 & y_1 & z_3 & z_4 \\ z_1^2 & y_2 & z_3^2 & z_4^2 \\ z_1^3 & y_3 & z_3^3 & z_4^3 \end{vmatrix}$$

$$= (z_1 - z_2)\begin{vmatrix} 0 & y_0 & 1 & 1 \\ 1 & y_1 & z_3 & z_4 \\ (z_1 + z_2) & y_2 & z_3^2 & z_4^2 \\ (z_1^2 + z_1z_2 + z_2^2) & y_3 & z_3^3 & z_4^3 \end{vmatrix}$$

and this is real since $(z_1 - z_2)$ is a pure imaginary and the minors of the real elements of the first column of the determinant are all pure imaginaries. Hence $A_1$ and $A_2$ are complex conjugates since when each is expressed as a quotient of two determinants by Cramer's rule, the sum of the two numerators is real and the common denominator is also real.

For purposes of numerical calculation $A_1$ may be obtained from

$$A_1 = \frac{NP}{D}$$

in which $D$ is obtained from VIIa,

$$N = y_3 - (z_2 + z_3 + z_4)y_2 + (z_2z_3 + z_2z_4 + z_3z_4)y_1 - (z_2z_3z_4)y_0,$$

$$\text{and } P = (z_2 - z_3)(z_2 - z_4)(z_3 - z_4)$$

$$= 2\beta[(\alpha - a)2b + \iota\{(\alpha - a)^2 + (\beta^2 - b^2)\}], \text{ a complex number.}$$

If $z_1z_2 = r^2$ and $z_3z_4 = \rho^2$, the symmetric functions of the $z$'s in the above formula may be calculated from

$$z_2z_3z_4 = (a - \iota b)\rho^2$$

$$z_2z_3 + z_2z_4 + z_3z_4 = \rho^2 + 2\alpha(a - \iota b)$$

$$z_2 + z_3 + z_4 = (a - \iota b) + 2\alpha$$

For the second case, which is exemplified by the vapor pressure data,

$$\text{(VIIb)} \qquad D = 2\iota b[(a - z_3)^2 + b^2]\,[(a - z_4)^2 + b^2]\,[z_3 - z_4],$$

a pure imaginary. The sum of the two numerators of $A_1$ and $A_2$, namely

$$(z_1 - z_2)\begin{vmatrix} 0 & y_0 & 1 & 1 \\ 1 & y_1 & z_3 & z_4 \\ z_1 + z_2 & y_2 & z_3^2 & z_4^2 \\ z_1^2 + z_1z_2 + z_2^2 & y_3 & z_3^3 & z_4^3 \end{vmatrix}$$

is a pure imaginary, since $(z_1 - z_2)$ has this character, and the determinant has nothing but real elements. Hence $A_1$ and $A_2$ are still complex conjugates when $z_3$ and $z_4$ are real, $z_1$ and $z_2$ being complex conjugates.

For purposes of numerical calculation $A_1$ may be obtained from

$$A_1 = \frac{N}{(z_1 - z_2)\,(z_1 - z_3)\,(z_1 - z_4)}.$$

Here $(z_1 - z_2)$ is a pure imaginary and the other three factors are complex.

Let

$$N = r_1(\cos\theta_1 + \iota\sin\theta_1)$$
$$z_1 - z_3 = r_2(\cos\theta_2 + \iota\sin\theta_2)$$
$$z_1 - z_4 = r_3(\cos\theta_3 + \iota\sin\theta_3)$$

Then

$$A_1 = \frac{r_1\,[\cos(\theta_1 - \theta_2 - \theta_3) + \iota\sin(\theta_1 - \theta_2 - \theta_3)]}{(z_1 - z_2)\,r_2\,r_3}$$

In calculating $N$ by the formula given for it in the preceding paragraph, the symmetric functions of the $z$'s were obtained from

$$z_2z_3z_4 = (a - \iota b)z_3z_4$$
$$z_2z_3 + z_2z_4 + z_3z_4 = (a - \iota b)(z_3 + z_4) + z_3z_4$$
$$z_2 + z_3 + z_4 = (a - \iota b) + z_3 + z_4.$$

## Example

The first two of the following tables are abstracted from Table 2, p. 178 of Bureau Standards Research Paper No. 523. The third table is abstracted from Table 3, p. 179 et. seq. of that publication. $x$ is the number of degrees centigrade *above* 100°. $y$ is the pressure of saturated water vapor in International Standard Atmospheres. In the first two of the following tables, the values of $y$ are observed values. In the third, they are interpolated or graduated values calculated at the Bureau of Standards.

TABLE I

| $x$ | $y$ |
|---|---|
| 0 | 1.0000 |
| 90 | 12.3887 |
| 180 | 63.3558 |
| 270 | 207.771 |

TABLE II

| $x$ | $y$ |
|---|---|
| 0 | 1.0000 |
| 50 | 4.6969 |
| 100 | 15.3472 |
| 150 | 39.2566 |
| 200 | 84.7969 |
| 250 | 163.205 |

TABLE III

| $x$ | $y$ |
|---|---|
| 0 | 1.0000 |
| 39 | 3.4666 |
| 78 | 9.4490 |
| 117 | 21.612 |
| 156 | 43.392 |
| 195 | 78.974 |
| 234 | 133.64 |
| 273 | 215.37 |

The observed values of $y$ in Table I are reproduced by the following formula used in conjunction with a standard six place table of logarithms and trigonometric functions:

$$\text{(I)} \qquad y = 3.967433\ \epsilon^{.01539540x} \cos(.4085758x - 75°24'03''.7).$$

The observed values of $y$ in Table II are reproduced by the following formula used in conjunction with a standard six place table of logarithms and trigonometric functions.

$$\text{(II)} \qquad \begin{aligned} y &= 3.0253744\ \epsilon^{.01515605x} \\ &+ 2.2171657\ \epsilon^{.011500716x} \cos(155°59'35''.5 - 0.7899232x). \end{aligned}$$

Hence the formula is presumably an excellent one for interpolation between the values of $y$ listed in Table II, if the greatest accuracy is not needed.[4]

The values of $y$ in Table III are reproduced exactly to five significant figures by the following formula used in conjunction with a standard six place table of logarithms and trigonometric functions.

$$y = 3.8902543\,\epsilon^{.01413920x} - .164787\,\epsilon^{-.0216930x}$$
$$+\,2.743000\,\epsilon^{.009884290x}\cos(.7860725x + 186°28'53''.2).$$

By means of this formula the saturation pressure of water vapor was calculated for every five degrees from 100°C to 370°C in order to make comparisons with the corresponding "smoothed" values in Table 2 of the Bureau of Standards publication referred to above. The discrepancies were never more than one in the fourth significant figure and generally less. The poorest agreement was in the ranges of temperature from 100°C to 135°C and from 245°C to 270°C.

It is a pleasure to acknowledge the intelligent and painstaking assistance of Mr. G. D. Lambert, undergraduate student at Washington University, for doing most of the computing.

Washington University,
St. Louis, Mo.

[4] The values of $y$ in Table III (not counting the value of $y$ for $x = 0$) are reproduced by it with an average error of .13% and a largest error (for $x = 234°$) of .30%. Four of the errors are negative and three positive.

# ON THE FREQUENCY DISTRIBUTION OF CERTAIN RATIOS

By H. L. Rietz

University of Iowa

Considerable interest in the distribution of ratios, $t = y/x$, has no doubt been suggested by important applications. For example, we may mention the opsonic index in bacteriology, the ratio of systolic to diastolic blood pressure in physiology, and ratios such as link relatives or certain index numbers in economics.

In 1910, Karl Pearson[1] gave certain properties of the distribution of ratios by means of approximate formulas for moments up to order four in terms of means, variances, product moments, and coefficients of variability of $x$ and $y$. The resulting formulas did not give, with sufficient accuracy, the constants of the distribution of the opsonic index for the purpose of Dr. Greenwood to whom Pearson attributed the derivation of the formulas for the special case in which $x$ and $y$ are uncorrelated. Pearson next adopted the plan of tabulating the reciprocals, say $x' = \frac{1}{x}$, and then finding the constants of the distribution of the product $yx'$ in the case in which $x'$ and $y$ are uncorrelated. He then obtained satisfactory results in illustrative examples.

In 1929, C. C. Craig[2] obtained the semi-invariants of $y/x$ in terms of moments of $x$ and $y$, and then expressed the moments in terms of the semi-invariants of the distribution function, $f(x, y)$, of $x$ and $y$. By this means, he was able to deal with the case in which $x$ and $y$ are normally correlated under suitable conditions. Craig found it desirable to restrict the distribution of $x$ in such a way that the probability of a zero value of $x$ is an infinitesimal of sufficiently high order that a certain integral exists. This limitation seems to imply in applications to actual data that no zero values of $x$ are to occur. This suggests that we deal with the cases of $x$ at or near zero with considerable care.

By starting with the assumption that the values of $x$ and $y$ are a set of normally distributed pairs of values with correlation coefficient $r$, and by considering the quotient $z = \frac{b + y}{a + x}$, $a$ and $b$ being constants, R. C. Geary,[3] in a paper published in 1930, found an algebraic function, $u = f(z)$, of fairly simple form with the property that $u$ is nearly normally distributed with arithmetic mean zero and standard deviation unity provided that $a + x$ is unlikely to

[1] On the constants of index distributions, Biometrika, Vol. 7 (1910), pp. 531–546.

[2] The frequency function of $y/x$, Annals of Mathematics, Vol. 30 (1928–29), pp. 471–486.

[3] The frequency distribution of the quotient of two normal variates, J. Royal Statistical Society, Vol. XCIII (1930), pp. 442–7.

have negative values. Here we have again a suggestion to exercise special care in the case of quotients with the divisor near zero or negative.

In 1932, Fieller[4] obtained in explicit form the approximate distribution of $t = y/x$ where values $(x, y)$ are drawn from the bivariate normal distribution

$$\frac{1}{2\pi\sigma_x\sigma_y\sqrt{1 - r^2}}\, e^{-\frac{1}{2}\frac{1}{1-r^2}\left\{\frac{(x-\bar{x})^2}{\sigma_x^2} + \frac{(y-\bar{y})^2}{\sigma_y^2} - 2r\frac{(x-\bar{x})(y-\bar{y})}{\sigma_x\sigma_y}\right\}}$$

under the condition that $\bar{x}$ is large compared with $\sigma_x$.

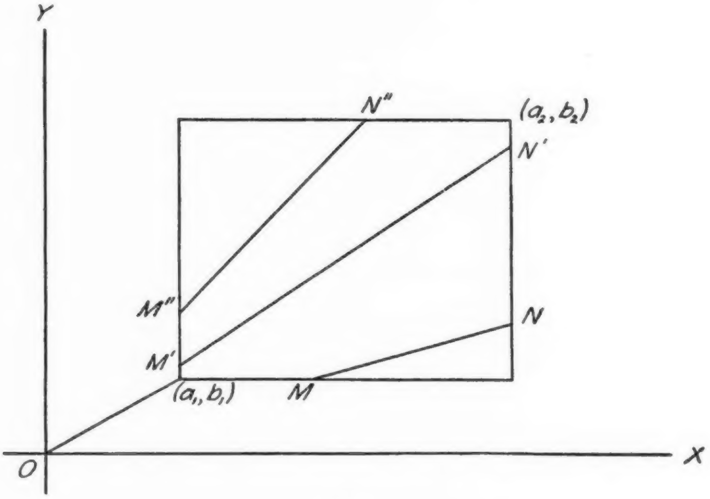

FIG. 1

Very recently Kullback[5] found the distribution law of the quotient, $t = y/x$, where $x$ and $y$ are drawn from Pearson Type III parent populations given by

$$f_1(x) = \frac{e^{-x}x^{p-1}}{\Gamma(p)}; \qquad f_2(y) = \frac{e^{-y}y^{q-1}}{\Gamma(q)}, \qquad 0 \leqq x \leqq \infty, \quad 0 \leqq y \leqq \infty.$$

It is fairly easy to see, in a general way, that the distribution of $t = y/x$ depends very much on the location of the origin as well as on the parent distribution from which $x$ and $y$ are drawn. This fact will be fairly obvious from the present paper whose main purpose is to give clear geometrical descriptions of the distributions of ratios, $t = y/x$, for each of several cases in which $(x, y)$ are points taken at random from certain simple geometrical figures conveniently located with respect to the origin.

In accord with the suggestions to be cautious when the divisor is near zero or negative, we consider first the very simple case of ratios $t = y/x$ obtained

[4] E. C. Fieller, The distribution of the index in a normal bivariate population, Biometrika, Vol. 24 (1932), pp. 428–440.

[5] Solomon Kullback, Annals of Mathematical Statistics, Vol. VII (1936), pp. 51–53.

from points uniformly distributed over a rectangle such as is shown in Fig. 1 with sides parallel to coordinate axes and $a_1 > 0$, $b_1 > 0$. As indicated on Fig. 1, we assume for simplicity that the coordinates of the points are positive and $a_1 \leqq x \leqq a_2$, $b_1 \leqq y \leqq b_2$.

Case I. When $\frac{b_1}{a_1} \leqq \frac{b_2}{a_2}$, Fig. 1.

Let $k\ dx\ dy$ be the probability that a point $(x, y)$ taken at random in the rectangle will fall into $dxdy$ where $k$ is a constant. Then

$$k \int_{b_1}^{b_2} \int_{a_1}^{a_2} dxdy = k(a_2 - a_1)(b_2 - b_1) = 1,$$

and

$$k = \frac{1}{(a_2 - a_1)(b_2 - b_1)}.$$

Transform the element $k\ dxdy$ into one with variables $t$, and $x$ by making

$$x = x,$$

$$y = tx.$$

The Jacobian is $|x| = x$.

The new element is $k\ x\ dxdt$ and is to be integrated over the range on $x$ for an assigned $t$ in order to get the probability, to within infinitesimals of higher order, that a random $t$ falls into an assigned $dt$. By assigning $t$ any value such that $\frac{b_1}{a_2} \leqq t \leqq \frac{b_1}{a_1}$, say $t$ is the slope of $MN$, (Fig. 1), we have

$$(1) \qquad k \int_{\frac{b_1}{t}}^{a_2} xdxdt = \frac{k}{2}\left(a_2^2 - \frac{b_1^2}{t^2}\right) dt$$

the limits of integration being indicated by the ends of the line $MN$.

When the assigned $t$ is such that $\frac{b_1}{a_1} \leqq t \leqq \frac{b_2}{a_2}$, say $t$ is the slope of the line $M'N'$, we have

$$(2) \qquad k \int_{a_1}^{a_2} x\,dx\,dt = \frac{k}{2}(a_2^2 - a_1^2)\,dt$$

When the assigned $t$ is such that $\frac{b_2}{a_2} \leqq t \leqq \frac{b_2}{a_1}$, say it is the slope of $M''N''$, we have

$$(3) \qquad k \int_{a_1}^{\frac{b_2}{t}} x\,dx\,dt = \frac{k}{2}\left(\frac{b_2^2}{t^2} - a_1^2\right) dt$$

Thus, from (1), (2), (3), when as in Fig. 1, $\frac{b_1}{a_1} \leqq \frac{b_2}{a_2}$, the frequency function of $t$ is given by

$$F(t) = \frac{k}{2}\left(a_2^2 - \frac{b_1^2}{t^2}\right) \quad \text{when} \quad \frac{b_1}{a_2} \leqq t \leqq \frac{b_1}{a_1}, \tag{4}$$

$$F(t) = \frac{k}{2}(a_2^2 - a_1^2) \quad \text{when} \quad \frac{b_1}{a_1} \leqq t \leqq \frac{b_2}{a_2}, \tag{5}$$

$$F(t) = \frac{k}{2}\left(\frac{b_2^2}{t^2} - a_1^2\right) \quad \text{when} \quad \frac{b_2}{a_2} \leqq t \leqq \frac{b_2}{a_1}. \tag{6}$$

See Fig. 2 for the general form of the frequency curve $F(t)$ when $\frac{b_1}{a_1} < \frac{b_2}{a_2}$ with the segment from $t = \frac{b_1}{a_1}$ to $\frac{b_2}{a_2}$ a horizontal straight line and with discontinuities in the first derivatives of $F(t)$ at $t = \frac{b_1}{a_1}$ and $t = \frac{b_2}{a_2}$.

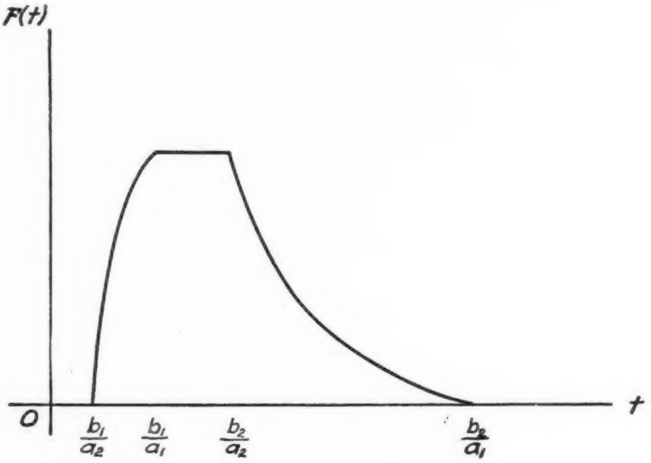

FIG. 2

When $a_1 \to 0$, and $b_1 = 0$, the frequency curve approaches

$$F(t) = \frac{a_2}{2b_2} \quad \text{when} \quad 0 \leqq t \leqq \frac{b_2}{a_2} \tag{7}$$

$$F(t) = \frac{b_2}{2a_2t^2} \quad \text{when} \quad t \geqq \frac{b_2}{a_2}. \tag{8}$$

It may be noted that the curve given by making $a_1 = 0$ and $b_1 = 0$ extends to infinity, and that the first and second moments about the origin are each infinite.

Case II. When $\frac{b_1}{a_1} > \frac{b_2}{a_2}$.

If the rectangle in Fig. 1 were moved upward keeping its sides parallel to the $x$ and $y$ axes until $\frac{b_1}{a_1} > \frac{b_2}{a_2}$, we would obtain

$$F(t) = \frac{k}{2}\left(a_2^2 - \frac{b_1^2}{t^2}\right) \quad \text{if} \quad \frac{b_1}{a_2} \leqq t \leqq \frac{b_2}{a_2}, \tag{9}$$

$$F(t) = \frac{k}{2t^2}\,(b_2^2 - b_1^2) \quad \text{if} \quad \frac{b_2}{a_2} \leqq t \leqq \frac{b_1}{a_1}, \tag{10}$$

$$F(t) = \frac{k}{2}\left(\frac{b_2^2}{t^2} - a_1^2\right) \quad \text{if} \quad \frac{b_1}{a_1} \leqq t \leqq \frac{b_2}{a_1}. \tag{11}$$

By comparing (5) and (10), it may be observed that $F(t)$ of the middle segment of the distribution curve differs much in Case II from its corresponding constant value in Case I.

By moving the rectangle of Fig. 1 downward, keeping its sides parallel to the $x$ and $y$ axes until $b_1$ is negative, we easily find further forms of the distribution curve $F(t)$.

To consider the distribution of the ratio $t = y/x$ for another very simple type of distribution of $x$ and $y$, suppose we have given the distribution function

$$f(x, y) = k\, e^{-\frac{x}{a} - \frac{y}{b}}, \begin{pmatrix} x \geqq c > 0,\ y \text{ non-negative} \\ a > c,\ b > 0 \end{pmatrix} \tag{12}$$

where $\int_0^\infty \int_c^\infty f(x, y)\, dx\, dy = 1$. Then

$$k = \frac{e^{c/a}}{ab}.$$

In this case,

$$\begin{aligned} F(t) &= \frac{e^{c/a}}{ab} \int_c^\infty x\, e^{-\frac{x}{a} - \frac{x\,t}{b}}\, dx \\ &= \frac{1}{b + at}\left(c + \frac{ab}{b + at}\right) e^{-\frac{c\,t}{b}}, \end{aligned} \tag{13}$$

a monotone decreasing function from $t = 0$ to $t = \infty$.

With $c = 0$ as a limiting value, we obtain

$$F(t) = \frac{ab}{(b + at)^2}, \tag{14}$$

a distribution curve with the mean value of $t$ at infinity.

If we should similarly consider

$$f(x, y) = \frac{2}{\pi \sigma_x \sigma_y} e^{-\frac{x^2}{2\sigma_x^2} - \frac{y^2}{2\sigma_y^2}} \text{ ($x$ and $y$ non-negative)} \tag{15}$$

we easily obtain

$$F(t) = \frac{1}{2\,\pi \sigma_x \sigma_y \left(\frac{1}{\sigma_x^2} + \frac{t^2}{\sigma_y^2}\right)} \tag{16}$$

as the distribution function.

Although the difficulties[6] of the problem of the distribution of the ratio $y/x$ when $x$ and $y$ are normally correlated have been overcome[7] to a considerable extent, still the examination of some very simple geometric cases of non-normal but linear correlation may not be without some interest. Such a case will now be considered.

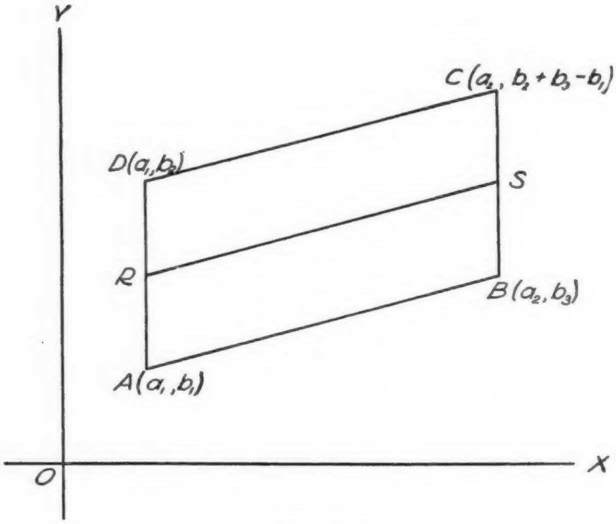

FIG. 3

For one very simple case in which $x$ and $y$ are correlated, suppose we are given a set of points $(x, y)$ uniformly distributed over the parallelogram $ABCD$ (Fig. 3) with sides $AD$ and $BC$ parallel to the $y$-axis so that the regression of $y$ on $x$ is linear as shown by the line $RS$.

The equation of $RS$ is

$$y = m(x - a_1) + \frac{b_1 + b_2}{2}. \tag{17}$$

[6] Loc. cit., Pearson, p. 531.

[7] Loc. cit., C. C. Craig, R. C. Geary, E. C. Fieller.

Then although $x_i$ and $y_i$ are correlated, $x_i$ and

$$y_i' = y_i - m(x_i - a_i) - \frac{b_1 + b_2}{2}$$

are uncorrelated. Let us consider the distribution of the ratio $t' = \frac{y_i'}{x_i}$.

Consider the element of frequency $kdxdy'$, where

$$k(b_2 - b_1)\,(a_2 - a_1) = 1. \tag{18}$$

Change variables to $x$ and $t'$ by the transformation

$$\begin{aligned} x &= x, \\ y' &= t'x. \end{aligned}$$

Then the element of frequency becomes

$$kx\;dx\;dt'. \tag{19}$$

Next integrate (19) with respect to $x$ under the restriction that $t'$ is assigned. Three cases occur:

(a) When $-\frac{b_2 - b_1}{2a_2} \leqq t' \leqq \frac{b_2 - b_1}{2a_2}$, we obtain by integration of (19) for the element of relative frequency of $t'$ in $dt'$,

$$k\int_{a_1}^{a_2} x\;dx\;dt' = \frac{k}{2}\,(a_2^2 - a_1^2)\;dt'. \tag{20}$$

(b) When $t' \geqq \frac{b_2 - b_1}{2a_2}$, we obtain

$$k\int_{a_1}^{\frac{b_2-b_1}{2t'}} x\;dx\;dt' = \frac{k}{2}\left[\frac{(b_2 - b_1)^2}{4t'^2} - a_1^2\right]dt' \tag{21}$$

(c) When $t' \leqq -\frac{b_2 - b_1}{2a_2}$, we similarly obtain

$$k\int_{a_1}^{-\frac{b_2-b_1}{2t'}} x\;dx\;dt' = \frac{k}{2}\left[\frac{(b_2 - b_1)^2}{4t'^2} - a_1^2\right]dt' \tag{22}$$

From (18), (19), (20), (21) and (22), the frequency function of $t'$ is given by

$$F(t') = \frac{a_2 + a_1}{2(b_2 - b_1)} \quad \text{when} \quad -\frac{b_2 - b_1}{2a_2} \leqq t' \leqq \frac{b_2 - b_1}{2a_2}; \tag{23}$$

$$F(t') = \frac{1}{2(b_2 - b_1)\,(a_2 - a_1)}\left[\frac{(b_2 - b_1)^2}{4t'^2} - a_1^2\right], \tag{24}$$

where the range of $t'$ is subject to either the inequalities,

$$\frac{b_2 - b_1}{2a_2} \leqq t' \leqq \frac{b_2 - b_1}{2a_1}, \quad \text{or} \quad -\frac{b_2 - b_1}{2a_1} \leqq t' \leqq -\frac{b_2 - b_1}{2a_2}.$$

See Fig. 4 for the general form of the $F(t')$ frequency curve. If we make $a_1 = 0$, the curve becomes infinite in range. If we make not only $a_1 = 0$, but $(b_1 + b_2)/2 = 0$, we have, in place of (17),

$$y = mx.$$

In this limiting situation, if we make $a_2 = a$ and $\frac{b_2 - b_1}{2} = b$,

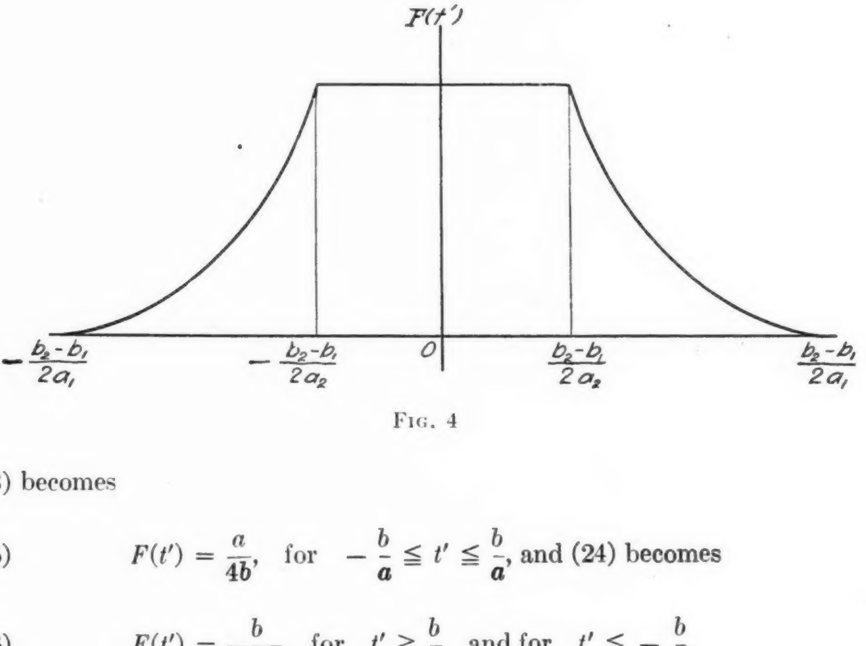

FIG. 4

(23) becomes

$$(25) \qquad F(t') = \frac{a}{4b}, \quad \text{for} \quad -\frac{b}{a} \leqq t' \leqq \frac{b}{a}, \text{ and (24) becomes}$$

$$(26) \qquad F(t') = \frac{b}{4\,at'^2} \quad \text{for} \quad t' \geqq \frac{b}{a} \quad \text{and for} \quad t' \leqq -\frac{b}{a}.$$

Then we have $y' = y - mx \cdot$

and $$t' = \frac{y'}{x} = t - m.$$

Further, if $t'$ is distributed in accord with a frequency function, $F(t')$, the distribution of $t = t' + m$ with $m$ constant is given by

$$F(t - m).$$

Hence, the probability that a random value $t$ will fall into a range $t$ to $t + dt$ is given to within infinitesimals of higher order by

$$\frac{a}{4b}\,dt \quad \text{when} \quad m - \frac{b}{a} \leqq t \leqq m + \frac{b}{a}, \tag{27}$$

and by

$$\frac{b\,dt}{4a(t - m)^2} \quad \text{when} \quad t \geqq m + \frac{b}{a} \text{ and } t \leqq m - \frac{b}{a}. \tag{28}$$

With the frequency curve given by (27) and (28) we may note that the variance of $t$ becomes infinite.

Without taking the space to continue illustrations, it is fairly obvious that a wide diversity of form can be given to the frequency function of the quotients $t = y/x$ by relatively simple changes in the location of a sample parent population with reference to the origin.

# EDITORIAL

## THE FUNDAMENTAL NATURE AND PROOF OF SHEPPARD'S ADJUSTMENTS

In the course of our discussion of moment adjustments, we shall have occasion to refer to the following lengthy distribution of discrete variates. By selecting

TABLE 1

*Distribution of the number of items correctly recorded by 244 students in a five minute code transcription test**

| Score $x$ | Freq. $f$ | Score $x$ | Freq. $f$ | Score $x$ | Freq. $f$ |
|---|---|---|---|---|---|
| 64 | 1 | 94 | 3 | 119 | 1 |
| 66 | 2 | 95 | 5 | 120 | 2 |
| 68 | 2 | 96 | 3 | 121 | 6 |
| 69 | 1 | 97 | 3 | 122 | 2 |
| 70 | 1 | 98 | 12 | 123 | 3 |
| 71 | 3 | 99 | 4 | 124 | 2 |
| 72 | 3 | 100 | 5 | 125 | 6 |
| 73 | 3 | 101 | 6 | 126 | 3 |
| 76 | 1 | 102 | 8 | 127 | 4 |
| 77 | 2 | 103 | 6 | 128 | 2 |
| 78 | 3 | 104 | 8 | 130 | 2 |
| 79 | 1 | 105 | 9 | 131 | 1 |
| 80 | 2 | 106 | 5 | 132 | 5 |
| 82 | 2 | 107 | 3 | 133 | 1 |
| 83 | 3 | 108 | 3 | 134 | 1 |
| 84 | 2 | 109 | 4 | 136 | 1 |
| 85 | 6 | 110 | 2 | 138 | 1 |
| 86 | 3 | 111 | 4 | 140 | 1 |
| 87 | 1 | 112 | 7 | 141 | 1 |
| 88 | 2 | 113 | 5 | 142 | 2 |
| 89 | 4 | 114 | 5 | 144 | 2 |
| 90 | 4 | 115 | 7 | 153 | 1 |
| 91 | 5 | 116 | 8 | 155 | 1 |
| 92 | 2 | 117 | 3 | | |
| 93 | 4 | 118 | 2 | Total | 244 |

* I am indebted to Professor J. A. Gengerelli, of the Department of Psychology of Univ. of California at Los Angeles, for these data.

the provisional mean, $M_0 = 105$, we find that

$$\Sigma\, x\, f = -129 \qquad \Sigma\, x^3 f = -52\ 005\ \ .$$

$$\Sigma\, x^2 f = 77\ 591 \qquad \Sigma\, x^4 f = 69\ 239\ 951.$$

Let us now form the nine possible distributions of grouped-discrete variates that arise from the nine possible "groupings of nine." These are presented in table 2.

TABLE 2

*Distributions derived from the data of table 1 by making the nine possible "groupings of nine"*

| First significant class interval of distribution | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| (1)<br>64–72 | (2)<br>63–71 | (3)<br>62–70 | (4)<br>61–69 | (5)<br>60–68 | (6)<br>59–67 | (7)<br>58–66 | (8)<br>57–65 | (9)<br>56–64 |
| 13 | 10 | 7 | 6 | 5 | 3 | 3 | 1 | 1 |
| 12 | 15 | 16 | 16 | 14 | 14 | 13 | 15 | 15 |
| 27 | 23 | 21 | 20 | 22 | 21 | 16 | 14 | 11 |
| 41 | 41 | 33 | 32 | 30 | 28 | 31 | 29 | 30 |
| 53 | 54 | 63 | 61 | 55 | 52 | 49 | 45 | 41 |
| 45 | 45 | 40 | 38 | 42 | 45 | 44 | 48 | 52 |
| 27 | 27 | 29 | 34 | 36 | 39 | 40 | 42 | 43 |
| 16 | 19 | 24 | 25 | 23 | 24 | 28 | 30 | 29 |
| 8 | 6 | 7 | 6 | 10 | 10 | 12 | 11 | 13 |
| 1 | 2 | 2 | 4 | 5 | 6 | 6 | 7 | 7 |
| 1 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 1 |
| | | | | | | | | 1 |

Let us now compute the values of $\Sigma xf$, $\Sigma x^2f$, $\Sigma x^3f$ and $\Sigma x^4f$ for each of the distributions of table 2, selecting $M_0 = 105$ in each instance in order to facilitate a comparison of these results with those for table 1. Thus, in spite of what would otherwise be called poor computing technique, we shall use the following class marks as values of $x$ for the first distribution above; $-37, -28, -19, \cdots, 35, 44, 53$. For the second we shall likewise use, $-38, -29, -20, \cdots, 34, 43, 52$, respectively.

TABLE 3

*Summations derived from the distributions listed in table 2, using* $M_0 = 105$

| Dist. | $\Sigma x f$ | $\Sigma x^2 f$ | $\Sigma x^3 f$ | $\Sigma x^4 f$ |
|---|---|---|---|---|
| (1) | − 181 | 77 149 | − 134 191 | 69 063 265 |
| (2) | − 218 | 78 466 | − 54 602 | 74 519 962 |
| (3) | − 111 | 77 769 | 2 889 | 71 465 409 |

TABLE 3—*Continued*

| Dist. | $\Sigma xf$ | $\Sigma x^2f$ | $\Sigma x^3f$ | $\Sigma x^4f$ |
|---|---|---|---|---|
| (4) | — 139 | 79 747 | — 23 311 | 74 171 443 |
| (5) | — 104 | 81 934 | 19 666 | 76 143 874 |
| (6) | — 87 | 80 145 | 16 551 | 72 467 541 |
| (7) | — 52 | 80 302 | — 36 118 | 71 851 930 |
| (8) | — 89 | 78 553 | — 101 357 | 68 426 497 |
| (9) | — 180 | 78 894 | — 180 792 | 73 155 150 |
| Average | — 129 | 79 217⅔ | — 54 585 | 72 362 785⅔ |

The fact that the average of the values of $\Sigma xf$ appearing in table 3 suggests that no adjustments of the first moment is necessary and that the variations in the nine values for $\Sigma xf$ may be regarded as *accidental errors* and attributed to grouping. An attempt to account for this phenomenon and also for the fact that the averages of the higher order summations of table 3 do not likewise agree with the corresponding summations of table 1 lead us directly to formulae for Sheppard's adjustments.

For the moment, let us concentrate our intention upon a single variate, $x_0$, and its associated frequency, $f_{x_0}$, that are a part of a distribution of discrete variates, such as table 1. Suppose we were to form the $k$ different distributions arising from the $k$ possible "groupings of $k$." In one of these distributions, $x_0$ will rest in the first position of a class interval: the limits of this class are $x_0$ and $(x_0 + k - 1)$ and the class mark is therefore $[x_0 + \frac{1}{2}(k - 1)]$. The contribution of the variate, $x_0$, to $\Sigma x^s f$ for this particular distribution is therefore

$$[x_0 + \tfrac{1}{2}(k - 1)]^s \cdot f_{x_0}.$$

If $x_0$ rests in the second position of a class, the limits of this class will be $(x_0 - 1)$ and $(x_0 + k - 2)$ and the corresponding class mark is $[x_0 + \frac{1}{2}(k - 3)]$ and the contribution of $x_0$ to $\Sigma x^s f$ for this distribution is

$$[x_0 + \tfrac{1}{2}(k - 3)]^s \cdot f_{x_0}.$$

The *expected* value of $\Sigma x^s f$ arising from the $k$ different groupings of variates is therefore,

$$(1) \qquad E(\sum x^s f) = \frac{1}{k}\Big[\sum^{1} x^s f + \sum^{2} x^s f + \cdots + \sum^{k} x^s f\Big]$$

where $\sum^{i} x^s f$ refers to that distribution in which a specified $x_0$ rests in the $i$-th position in the class in which it occurs. The contribution of $x_0$ to this expected value is therefore

$$(2) \quad \frac{1}{k}\{[x_0 + \tfrac{1}{2}(k-1)]^s + [x_0 + \tfrac{1}{2}(k-3)]^s + [x_0 + \tfrac{1}{2}(k-5)]^s + \cdots\} f_{x_0},$$

this series consisting obviously of $k$ terms.

Expanding each term of (2) by the binomial theorem yields

$$\frac{1}{k}\left[x_0^s - {}_sC_1\, x_0^{s-1}\left(\frac{k-1}{2}\right) + {}_sC_2\, x_0^{s-2}\left(\frac{k-1}{2}\right)^2 - {}_sC_3\, x_0^{s-3}\left(\frac{k-1}{2}\right)^3 + \cdots\right]$$

$$\frac{1}{k}\left[x_0^s - {}_sC_1\, x_0^{s-1}\left(\frac{k-3}{2}\right) + {}_sC_2\, x_0^{s-2}\left(\frac{k-3}{2}\right)^2 - {}_sC_3\, x_0^{s-3}\left(\frac{k-3}{2}\right)^3 + \cdots\right]$$

$$\frac{1}{k}\left[x_0^s - {}_sC_1\, x_0^{s-1}\left(\frac{k-5}{2}\right) + {}_sC_2\, x_0^{s-2}\left(\frac{k-5}{2}\right)^2 - {}_sC_3\, x_0^{s-3}\left(\frac{k-5}{2}\right)^3 + \cdots\right]$$

etc.

Since $s$ is an integer, series (2) may be written as the sum of the $(s + 1)$ terms of the series

$$(3) \quad [x_0^s\, S_0 - {}_sC_1\, x_0^{s-1}\, S_1 + {}_sC_2\, x_0^{s-2}\, S_2 - {}_sC_3\, x_0^{s-3}\, S_3 + \cdots] f_{x_0},$$

where

$$S_i = \frac{1}{k}\left[\left(\frac{k-1}{2}\right)^i + \left(\frac{k-3}{2}\right)^i + \left(\frac{k-5}{2}\right)^i + \cdots \text{ to } k \text{ terms}\right].$$

By the Euler-Maclaurin Sum Formula we have

$$\sum_{x=a}^{b} x^p = \frac{1}{p+1}(b^{p+1} - a^{p+1}) + \tfrac{1}{2}(b^p + a^p) + \frac{B_1}{2!}p\,(b^{p-1} - a^{p-1})$$
$$- \frac{B_3}{4!}p^{(3)}(b^{p-3} - a^{p-3}) + \frac{B_5}{6!}p^{(5)}(b^{p-5} - a^{p-5}) + \cdots,$$

where $p^{(i)} = p(p-1)(p-2)(p-3)\cdots$ to $i$ factors. In our expression for $S_i$, $a = \frac{1}{2}(k-1) = -b$, and therefore $S_i$ equals zero when $i$ is an odd integer. For even values of $i$,

$$(4) \quad S_i = \frac{2}{k}\left\{\frac{(k-1)^i(k+i)}{2^{i+1}(1+i)} + \frac{B_1}{2!}\,i\left(\frac{k-1}{2}\right)^{i-1} - \frac{B_3}{4!}\,i^{(3)}\left(\frac{k-1}{2}\right)^{i-3} + \frac{B_5}{6!}\,i^{(5)}\left(\frac{k-1}{2}\right)^{i-5} - \cdots\right\}$$

so that

$$S_0 = 1$$

$$S_2 = \frac{1}{12}(k^2 - 1)$$

$$S_4 = \frac{1}{240}(k^2 - 1)(3k^2 - 7)$$

$$S_6 = \frac{1}{1344}(k^2 - 1)(3k^4 - 18k^2 + 31)$$

etc.

Since expression (3) represents the contribution of any variate, $x_0$, to the expected value defined by (1), we may obtain by summation

$$(5) \quad E(\sum x^s f) = \sum x^s f + {}_sC_2 \cdot S_2 \cdot \sum x^{s-2} f + {}_sC_4 \cdot S_4 \sum x^{s-4} f + \cdots .$$

To illustrate: if we desire to shorten the distribution of table 1 by forming class intervals of dimension 9,

$$S_2 = \frac{1}{12}(9^2 - 1) = \frac{20}{3}, \qquad S_4 = \frac{1}{240}(9^2 - 1)(3 \cdot 9^2 - 7) = \frac{236}{3},$$

and by formula (5),

$$E(\sum x\ f) = \sum x f = -129$$

$$E(\sum x^2 f) = \sum x^2 f + {}_2C_2 \cdot S_2 \cdot \sum f = 77591 + \frac{20}{3} \cdot 244 = 79217^{2/3}$$

$$E(\sum x^3 f) = \sum x^3 f + {}_3C_2 \cdot S_2 \cdot \sum x f = -52005 + 3 \cdot \frac{20}{3}(-129) = -54585$$

$$E(\sum x^4 f) = \sum x^4 f + {}_4C_2 \cdot S_2 \cdot \sum x^2 f + {}_4C_4 \cdot S_4 \cdot \sum f$$

$$= 69239951 + 6 \cdot \frac{20}{3} \cdot 77591 + \frac{236}{3} \cdot 244 = 72362785^{2/3}.$$

Since these expected values are identical with those computed directly in table 3, we see that formula (5) provides the adjustments necessary to eliminate the effect of the systematic errors caused by grouping.

Dividing both sides of (5) by $\Sigma f$ yields

$$(6) \quad E(\mu'_s) = \mu'_s + {}_sC_2 \cdot S_2 \cdot \mu'_{s-2} + {}_sC_4 \cdot S_4\, \mu'_{s-4} + {}_sC_6 \cdot S_6 \cdot \mu'_{s-6} + \cdots ,$$

that is

$$E(\mu'_1) = \mu'_1$$

$$E(\mu'_2) = \mu'_2 + \frac{1}{12}(k^2 - 1)$$

$$E(\mu_3') = \mu_3' + \frac{3}{12}(k^2 - 1)\,\mu_1'$$

$$E(\mu_4') = \mu_4' + \frac{6}{12}(k^2 - 1)\,\mu_2' + \frac{1}{240}(k^2 - 1)(3k^2 - 7)$$

$$E(\mu_5') = \mu_5' + \frac{10}{12}(k^2 - 1)\,\mu_3' + \frac{5}{240}(k^2 - 1)(3k^2 - 7)\,\mu_1'$$

$$\cdot \qquad \text{etc.}$$

In numerical computations we generally prefer to select the class interval as the unit of $x$ and in this case we have

$$E(\mu_1') = \mu_1'$$

$$E(\mu_2') = \mu_2' + \frac{1}{12}\left(1 - \frac{1}{k^2}\right)$$

$$E(\mu_3') = \mu_3' + \frac{3}{12}\left(1 - \frac{1}{k^2}\right)\mu_1'$$

$$E(\mu_4') = \mu_4' + \frac{6}{12}\left(1 - \frac{1}{k^2}\right)\mu_2' + \frac{1}{240}\left(1 - \frac{1}{k^2}\right)\left(3 - \frac{7}{k^2}\right)$$

$$\text{etc.}$$

Ordinarily we are interested in estimating the values of the moments that would have been obtained if we had not used the time-saving device of grouping the variates and therefore we solve the previous set of equations for the moments of the ungrouped distribution and obtain

$$(7) \qquad \left\{ \begin{aligned} \mu_1' &= E(\mu_1') \\ \mu_2' &= E(\mu_2') - \frac{1}{12}\left(1 - \frac{1}{k^2}\right) \\ \mu_3' &= E(\mu_3') - \frac{3}{12}\left(1 - \frac{1}{k^2}\right)E(\mu_1') \\ \mu_4' &= E(\mu_4') - \frac{6}{12}\left(1 - \frac{1}{k^2}\right)E(\mu_2') + \frac{1}{240}\left(1 - \frac{1}{k^2}\right)\left(7 - \frac{3}{k^2}\right) \\ &\qquad \text{etc.} \end{aligned} \right.$$

In general we may write, corresponding to formula (6),

$$(8) \qquad \mu_s' = E(\mu_s') - {}_sC_2 \cdot P_2 \cdot E(\mu_{s-2}') + {}_sC_4 \cdot P_4 \cdot E(\mu_{s-4}') - \cdots$$

where

$$P_2 = \frac{1}{12}\left(1 - \frac{1}{k^2}\right)$$

$$P_4 = \frac{1}{240}\left(1 - \frac{1}{k^2}\right)\left(7 - \frac{3}{k^2}\right)$$

$$P_6 = \frac{1}{1344}\left(1 - \frac{1}{k^2}\right)\left(31 - \frac{18}{k^2} + \frac{3}{k^4}\right)$$

$$P_8 = \frac{1}{11520}\left(1 - \frac{1}{k^2}\right)\left(381 - \frac{239}{k^2} + \frac{55}{k^4} - \frac{5}{k^6}\right)$$

$$P_{10} = \frac{1}{33792}\left(1 - \frac{1}{k^2}\right)\left(2555 - \frac{1636}{k^2} + \frac{410}{k^4} - \frac{52}{k^6} + \frac{3}{k^6}\right)$$

$$P_{12} = \frac{1}{5591040}\left(1 - \frac{1}{k^2}\right)\left(1414477 - \frac{910573}{k^2} + \frac{233570}{k^4} - \frac{32410}{k^6} + \frac{2625}{k^8} - \frac{105}{k^{10}}\right).$$

In actual problems we do not know the exact values of the expectations involved in formulae (7) and (8), and are forced to obtain mere approximations by utilizing in their stead the corresponding moments computed from the single chance grouped distribution. These approximations correspond to those employed in the theory of probable error, namely, substitutions of the moments derived from a single sample for the corresponding expected moments of the parent population.

The adjustments so far considered may properly be referred to as *Sheppard's adjustments about a fixed point.* At first thought it might appear that we might obtain corresponding formulae for the expectations of moments *about the mean* by merely dropping the primes in formula (6) and obtain, for example,

$$\mu_2 = E(\mu_2) - \frac{1}{12}(k^2 - 1),$$

but unfortunately this is not true. For example, the exact value for the variance of the distribution of table 1 is 18915563/244². Using the summations of table 3 and computing the variance for each of the nine groupings yields

$$\begin{aligned} E(\mu_2) &= \frac{1}{9.244^2}[18791595 + 19098180 + 18963315 + 19438947 \\ &\quad + 19981080 + 19547811 + 19590984 + 19159011 + 19217736] \\ &= 19309851/244^2. \end{aligned} \tag{9}$$

Since $\frac{1}{12}(k^2 - 1) = \frac{1}{12}(9^2 - 1) = 20/3$ we see that

$$\mu_2 < E(\mu_2) - \frac{1}{12}(k^2 - 1).$$

In the theory of sampling we differentiate between the standard errors of moments about a fixed point and the standard error of moments about the mean of the sample. Apparently writers on the subject of Sheppard's adjustments have overlooked the case of adjustments about the mean, although the solution for the second moment is readily obtained as follows:

$$\begin{aligned} E(\mu_2) &= E(\mu_2' - M^2) = E(\mu_2') - E(M^2) \\ &= \mu_2' + \frac{1}{12}(k^2 - 1) - \frac{1}{k}(M_1^2 + M_2^2 + \cdots + M_k^2), \end{aligned}$$

where $M_i$ represents the mean of the $i$-th of the $k$ different grouped distributions. Since

$$\mu_2 = \mu_2' - M^2 = \mu_2' - \frac{1}{k}(M_1 + M_2 + \cdots + M_k),$$

$$\begin{aligned} E(\mu_2) = \mu_2 &+ \frac{1}{12}(k^2 - 1) \\ &- \left[\frac{M_1^2 + M_2^2 + \cdots + M_k^2}{k} - \left(\frac{M_1 + M_2 + \cdots + M_k}{k}\right)^2\right]. \end{aligned}$$

But since for any set of $k$ variates

$$\sigma_\nu^2 = \frac{\Sigma\nu^2}{k} - \left(\frac{\Sigma\nu}{k}\right)^2,$$

we have that

$$E(\mu_2) = \mu_2 + \frac{1}{12}(k^2 - 1) - \sigma_M^2. \tag{10}$$

Referring back to table 3 we find that

$$\sigma_M^2 = \frac{7856}{3.(244^2)}$$

and the numerical results now satisfy equation (10).

For the benefit of those interested in unsolved problems of mathematical statistics we may say that nothing appears to have been written as yet on the most important problem associated with the systematic errors due to grouping. It is of course desirable to eliminate these systematic errors introduced by grouping, but it is even more important to investigate the distribution of the accidental errors that remain after the systematic errors have been eliminated. For example it is gratifying to know that no systematic errors are present in the $\Sigma xf$ column of table 3 and that equation (6) will enable us to add a constant to each summation of the $\Sigma x^3f$ column so that the mean of these adjusted values will agree with the value $\Sigma x^3f = -52005$ obtained in table 1. It is rather disconcerting, however, to realize that in actual practice we *may* in the case of discrete variates and *must* in the case of continuous variates select an arbitrary set of class limits for our recorded data, and that after adjustments for grouping have been made, our estimates of the true values of the moments of the distribution will—as in table 3—depend so much upon the choice of these limits. Thus, the standard error of the mean attributed to grouping is

$$\sigma_M = \frac{1}{244}\sqrt{\frac{7856}{3}} = 0.21\,,$$

which is about twenty percent as large as the approximation for the standard error of the mean due to sampling from an infinite parent population, namely,

$$\sigma_M = \frac{\sigma}{\sqrt{N}} = 1.15\,.$$

If one will take the trouble to compute the values of $\mu_3$ and $\mu_4$ for each of the distributions of table 2, utilizing the summations of table 3, and then compute and compare the values of $\sigma_{\mu_3}$ and $\sigma_{\mu_4}$ due to grouping with the corresponding functions associated with sampling, he will realize the seriousness of the situation.

## Summary

The formula for Sheppard's adjustments for distributions of grouped discrete variates was first given without proof in the Editorial of Vol. 1, No. 1 of the *Annals* (page 111). The method used to develop the general formula was extremely laborious and paralleled the method used for the case of continuous variates in the *Handbook of Mathematical Statistics*, Chapter 7, except that the calculus of finite differences was employed. A more satisfactory proof of this formula was presented by Dr. J. R. Abernethy in Vol. 4, No. 4 of the *Annals* in an article entitled "*On the Elimination of Systematic Errors Due to Grouping.*" An extremely elegant development of the same formula and an extension to the case of two variables appears elsewhere in this volume by Professor C. C. Craig. From the point of view of expectations, all of these developments are adjustments about a fixed point, although this fixed point may be selected arbitrarily at the mean of the distribution in question. The obtaining of formulae for the adjustments about the mean of each grouping and the distribution of the accidental errors that remain after these systematic errors have been removed has apparently been neglected to date and should interest students of mathematical statistics.

From a mathematical standpoint, the development of this paper is the simplest of all that have appearéd to date: the adjustments for the first four moments can be worked out with the aid of the binomial considerations leading to formula (3) and the following well known formulae for the sums of the powers of the first $n$ integers:

$$S_1 = \frac{n(n+1)}{2} \qquad S_3 = \frac{n^2(n+1)^2}{4}$$

$$S_2 = \frac{n(n+1)(2n+1)}{6} \qquad S_4 = \frac{n(n+1)(2n+1)(3n^2+3n-1)}{30}.$$

One should note that the condition of high contact is not required in this paper or in the developments of Abernethy or Craig. The results of the three preceding papers agree with those obtained about a fixed point in this paper, but fail to hold for the case of expectations about the mean, if we accept the following definition:

$$E(\mu_s) = \frac{1}{k}(\mu_{s:1} + \mu_{s:2} + \cdots + \mu_{s:k}), \qquad (s = 2, 3, \cdots)$$

where $\mu_{s:i}$ designates the $s$-th moment computed about the mean of the $i$-th grouped distribution, $(1 \leqq i \leqq k)$.

H. C. CARVER.